LASSO ASYMPTOTICS FOR HEAVY TAILED ERRORS

A Dissertation

by

AARON SETH GOLDSMITH

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Joel Zinn
Committee Members,	Ursula Müller
	Maurice Rojas
	Thomas Schlumprecht
Head of Department,	Emil Straube

December 2015

Major Subject: Mathematics

ABSTRACT

We consider the asymptotic behavior of the ℓ^1 regularized least squares estimator (LASSO) for the linear regression model

$$Y = X\beta + \xi$$

with training data $(X, Y) \in \mathbb{R}^{n\times p} \times \mathbb{R}^n$, true parameter $\beta \in \mathbb{R}^p$, and observation noise $\xi \in \mathbb{R}^n$. The LASSO estimator, defined by

$$\hat{\beta}_n :\in \operatorname*{arg\,min}_{u\in\mathbb{R}^p} \left[\|Xu - Y\|_2^2 + \lambda_n \|u\|_1\right],$$

introduces a bias toward 0 to encourage sparse estimates. LASSO has become a staple in the statistician's breadbasket; it behaves very well and is quickly computed.

In the case that ξ_i are i.i.d. with $\mathbb{E}|\xi_i|^\alpha < \infty$ for some $1 < \alpha < 2$, Chatterjee and Lahiri found the exact rate, almost surely, for which the LASSO $\hat{\beta}_n$ tends to β. We consider instead ξ_i that are i.i.d., possess all moments less than α, and eventually nearly follow a Pareto tail $\mathbb{P}\{|\xi_i| > t\} \approx t^{-\alpha}$. Specifically, we only require the tails of ξ_i to be regularly varying.

We center and scale both the quantity inside the arg min and $\hat{\beta}_n$ itself to prepare for a CLT. We find conditions that promise both convergence (uniformly over a class of designs $\mathfrak{X}$) of the quantity inside the arg min and uniform tightness of the centered, scaled $\hat{\beta}_n$. Then, we use a standard theorem to pass to uniform convergence of the centered, scaled $\hat{\beta}_n$. Finally, we use a basic inequality to prove rate consistency for $\hat{\beta}_n$ when p is allowed to increase with n.

DEDICATION

I dedicate this work to my grandfather, Fred Cochran, who led an exemplary life. He was an amalgamation of kindness, patience, and diligence. After I wrecked myself in a providential car accident, he took care of me and allowed me to continuously work Putnam problems in my bedridden state. I now cherish those weeks in which I had the opportunity to observe both my grandparents' daily living.

It was his pride in my accomplishments that gave me reason to persist. He couldn't seem to wait to call me doctor, and made grand gestures to attend my graduation, despite the toll his cancer exacted. He was eventually able to call me doctor shortly after my defense, over the phone.

Unfortunately, he is now gone, and our memories of him are complete. At the same time, the good memories will never fade.

ACKNOWLEDGEMENTS

I owe thanks to many people; this page is insufficient. As a matter of fact, it is also unnecessary. Yet, it is well deserved.

Of course, my parents have always supported my hunger for mathematics. Dad gave me his old CRC handbook called Standard Mathematical Tables, introducing me to Taylor series. Unforgettable. Thank you. They also performed duties ranging from competition travels starting on Saturdays as early as 4 AM to trying to understand my incomplete thoughts on convex geometry.

Annie, my wife, has valiantly endured the general mental absence that comes to me when working on hard problems. She also listens to my incomplete thoughts about many things. She doesn't realize her virtues. It has been a privilege to see her every day.

Finally, the Texas A&M Math Department has generously assisted me financially every semester and has been a delightful place to work, along with outstanding characters. Joel Zinn has shown me bottomless patience, which I appreciate more than I let on. He is rarely too busy to meet, despite his penchant for taking on projects. Monique Stewart works very hard shepherding us, making sure our skies are clear and ready for takeoff. Peter Howard always has his eyes looking out for our best interest, doing whatever it takes to benefit our graduate program. And, Parth Sarin persevered with me until we generated images to my satisfaction and laughed along the way. Thank you all very much.

NOMENCLATURE

α	index of stability $\in (1, 2)$
ξ	α-regularly varying error
$r(t), R(t)$	regularly varying tails (left, right) of ξ
$\mathcal{R}^\alpha(t)$	$r^\alpha(t) + R^\alpha(t)$
b_n	$\mathcal{R}^{-1}(n)$
X	$n \times p$ data matrix
Y	response
λ	tuning constant
$\hat{\beta}_n$	LASSO solution
$\tilde{\beta}_n$	OLS solution
$\hat{u}_n$	$b_n(\hat{\beta}_n - \beta)$
u_S	$u\vert_{\mathbb{R}^S}$
Φ	Lèvy measure on $\mathbb{R}^p$
φ	spectral measure of Φ (defined on on $\mathbb{S}^{p-1}$)
$\vert\Phi\vert, \vert\varphi\vert$	total variation, e.g. $\Phi(\mathbb{R}^p), \varphi(\mathbb{S}^{p-1})$
ch.f.	characteristic function operator
$*$	convolution operator
$\rightsquigarrow$	weak convergence
$\sim$	equal in distribution
$A \succ \mathrm{B}$	$A - B$ is positive definite
$\overline{\lim}$	$\limsup_n$

TABLE OF CONTENTS

Page

ABSTRACT ii

DEDICATION iii

ACKNOWLEDGEMENTS iv

NOMENCLATURE v

TABLE OF CONTENTS vi

1. INTRODUCTION 1

1.1 Stable Distributions and Regular Variation 2
1.2 Sparsity and Regularization 6
1.3 Noisy Problem 9
1.4 Equivalence of Limits of Sums 10
1.5 Centering and Scaling the LASSO 13
1.6 The Argmin Theorem 17

2. FIXED DESIGN, FIXED NUMBER OF REGRESSORS 19

2.1 Laws of Large Numbers 19
2.2 Infinitely Divisible Central Limit Theorem in $\mathbb{R}^p$ 23
2.3 Applications to LASSO 32

3. VARIABLE DESIGN, FIXED NUMBER OF REGRESSORS 42

3.1 A General CLT 42
3.2 Choosing a Semimetric 44
3.3 Convergence of the Cross Term 45
3.4 LASSO 49

4. STOCHASTIC BOUNDEDNESS WITH INCREASING NUMBER OF REGRESSORS 55

4.1 Boundedness of Z_n 55
4.2 The Oracle 58

5. INCREASING NUMBER OF REGRESSORS 62

5.1 Reformulation of $\sqrt{\text{LASSO}}$ 63
5.2 Controlling the Cross Term 65
5.3 Subregressions 68

6. CONCLUSIONS 69

REFERENCES 70

1. INTRODUCTION

We study the linear regression model $Y = X\beta+\xi$, where we are handed $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^n$, and then asked to learn about $\beta \in \mathbb{R}^p$, under some model assumptions on the observation noise $\xi \in \mathbb{R}^n$. Generally, this is contingent on two properties.

(1) The design X must see β well enough so that $X\beta$ does not lose knowledge of β. Particularly, we must have $X\beta \neq 0$.

(2) The random noise ξ must not overcorrupt the data.

Neither property is checkable in practice, but this is the nature of the beast in statistics. Our goal is to mitigate this inconvenience by finding conditions on X that capture (1) and (2) with the fewest assumptions on β and ξ. We are only concerned with asymptotics, so we do not necessarily seek sharp estimates, and our conditions only need to hold eventually.

Ordinary least squares (OLS) is ubiquitous in statistical applications. It's long-standing (Gauss 1795, Legendre 1805), easily implemented, efficiently computed, and has a simple interpretable solution. Unfortunately, OLS is misapplied for various reasons, including the fact that practitioners are often not aware of alternatives. Drawbacks to OLS include nonuniquenss, nonrobustness, and overfitting. A popular fix is to assume the true value β has a specific structure, then force the estimate to have the same structure, whence come regularized estimators like ridge regression and LASSO (see Section 1.2).

For observation noise ξ_i with an α-moment ($\alpha \in (1, 2)$), Chatterjee and Lahiri

[10] found the exact rate of the LASSO

$$\hat{\beta}_n :\in \operatorname*{arg\,min}_{u \in \mathbb{R}^p} \left[\|Xu - y\|_2^2 + \lambda \|u\|_1 \right]$$

to be $\hat{\beta}_n - \beta = O(n^{-1/\alpha})$ almost surely. We will simplify their proof (Remark 2.3.6). However, we are primarily concerned with distributional statements, wherein regularly varying noise of order α is more fitting than noise with an α-moment [Fe, BGT]. Then, in the i.i.d. central limit theorem for regularly varying distributions, the normal distribution is replaced by the family of stable distributions. For background on stable distributions and regular variation, see Section 1.1.

Section 2 works the finite dimensional case. Lemma 2.3.5 shows that $\hat{\beta}_n$ has the same rate when ξ_i is regularly varying of order α as the rate found in [10] when ξ_i has an α-moment. Theorem 2.3.7 goes further to a CLT, then sections 4 and 3 generalize to two different frameworks: when convergence is uniform over a class of designs $\mathfrak{X}$, and when the number of regressors p is allowed to increase with n, respectively. Theorem 4.2 from [1] will be centrally important to both these endeavors. Our Theorems 3.4.4 and 4.2.1 are our main results.

1.1 Stable Distributions and Regular Variation

Chatterjee and Lahiri used a weighted version of the Markinciewicz strong law of large numbers to get the exact almost sure rate of the LASSO estimator when the errors ξ_i possess an α-moment [10]. It is natural, therefore, to try to prove the same exact rate using a weak law of large numbers. This takes place in Section 2.1.

The central limit theorem prototype is "the sum of many small independent random quantities is approximately normally distributed," and so the normal distribution is most important in statistics. Still, there are other possible limit distributions that fit this prototype, even if the summands are required to be identical. Moreover,

there are many data sets which do not appear to settle into a normal distribution, but rather into a distribution from a more general class called *stable* distributions. The first instance known to the author is Mandelbrot's long tail suspicion about *certain* financial data [21], when distributions had uncommonly distant outliers.

Definition 1.1.1. *Call a distribution F infinitely divisible if for any n, there is a distribution G such that F is the n-fold convolution $G^{n*} = F$. Call a distribution F stable if there are location parameters a_r, b_r $(r > 0)$ so that $F^{r*} = F(a_r t + b_r)$.*

The characteristic function is a convenient tool, especially for distributions with no closed form. Kolmogorov (see [18] Ch. 18) found the canonical representation of the characteristic function of a general infinitely divisible distribution. We will use the multivariate version in Theorem 3.1.11 from [22] (pg. 41).

Theorem 1.1.2 (Lévy Representation)**.** *The Lévy representation of the log characteristic function of an infinitely divisible random vector $Z \in \mathbb{R}^p$ is*

$$\log(\text{ch.f. } Z) = i\langle a, u\rangle - Q(u) + \int_{x\neq 0} \left(e^{i\langle x,u\rangle} - 1 - i\langle x, u\rangle \mathbb{1}\{\|x\| \leq 1\}\right) d\Phi(x) \quad (1.1)$$

where the centering $a \in \mathbb{R}^p$, the normal component $Q(u)$ is a semidefinite quadratic form ($Q(u) = \langle u, \text{Cov}(Z)u\rangle$), and the Lévy measure Φ is a Borel measure on $\mathbb{R}^p \setminus \{0\}$ satisfying

$$\int_{x\neq 0} \min(\|x\|^2, 1) d\Phi(x) < \infty$$

The triple $[a, Q, \Phi]$ uniquely determines the law of Z.

Remark 1.1.3. *With a change of centering a, the $i\langle u, x\rangle \mathbb{1}\{\|x\| \leq 1\}$ term in the integrand in Theorem 1.1.2 could be replaced with $i\langle u, x\rangle/(1 + \|x\|^2)$ or any other bounded function that behaves like $i\langle u, x\rangle$ near $x = 0$. See [18] for more discussion.*

If we equip $\mathbb{R}^p$ with spherical coordinates and if there are measures μ, φ on $\mathbb{R}_+, \mathbb{S}^{p-1}$ (resp.) and $\alpha \in (0,2)$ such that

$$\begin{aligned} d\Phi(r,\theta) =& d\mu(r)d\varphi(\theta) \\ \mu[r,\infty) =& r^{-\alpha} \end{aligned}$$

then $Z \sim [0,0,\Phi]$ is stable as in Definition 1.1.1 for $a_r = r^{1/\alpha}$ and $b_r = 0$ (Theorem 7.3.3 on pg. 263 of [22]). For centered stable distributions on $\mathbb{R}$ with $\alpha \neq 1$, this means i.i.d. sums of copies of Z only differ from Z by a scale parameter (See [13]):

$$r^{1/\alpha}Z^1 + s^{1/\alpha}Z^2 \sim (r+s)^{1/\alpha}Z \tag{1.2}$$

where Z^1, Z^2 are i.i.d copies of Z and $r, s \geq 0$. Note, if Z is normal ($\alpha = 2$), this is simply the additive property of variance, even though (1.2) did not technically address $\alpha = 2$.

The rest of this section can be found in [7] and has to do with a concept intimately connected to stable distributions, namely regular variation. A distribution has regularly varying tails iff it satisfies an i.i.d. central limit theorem, converging to a stable distribution (see [15, 25]).

Definition 1.1.4. *Call $R(t)$ regularly varying at infinity if $\lim_{t\to\infty} \frac{R(ct)}{R(t)}$ exists for each $c > 0$. Call $R(t)$ slowly varying at infinity if the limit is 1, regardless of c.*

Surprisingly, the only possible limits are powers, as seen in the Uniform Convergence Theorem (pg. 275 of [15] or pg. 22 of [7]).

Theorem 1.1.5 (Uniform Convergence Theorem)**.** *If $R(t)$ is regularly varying at*

infinity, there is a number $\alpha \in \mathbb{R}$ *so that for each* $c > 0$,

$$\lim_{t\to\infty} \frac{r(ct)}{r(t)} = c^{\alpha}$$

Moreover, this convergence is uniform over c *in*

$$\begin{cases} [a, \infty) & \text{if } \alpha < 0 \\ [a, b] & \text{if } \alpha = 0 \\ (0, b] & \text{if } \alpha > 0 \end{cases}$$

Definition 1.1.6. *Call such* $R(t)$ *in the previous theorem regularly varying of order* α.

Theorem 1.1.7. *If* $R(t)$ *is regularly varying of order* α, *then* $t^{-\alpha}R(t)$ *is slowly varying.*

Furthermore, slowly varying functions have the following characterization in what is called Karamata's Representation Theorem (pg. 12 of [7]):

Theorem 1.1.8 (Representation Theorem)**.** *The function* ℓ *is slowly varying iff it may be written in the form*

$$\ell(t) = c(t)\exp\left(\int_a^t \epsilon(u)\frac{du}{u}\right)$$

for some $a > 0$, *where* $c(\cdot)$ *is measureable,* $c(t) \to c \in (0, \infty)$ *and* $\epsilon(t) \to 0$ *as* $t \to \infty$.

Theorem 1.1.9. *Every slowly varying function*

$$\ell(t) = c(t)\exp\left(\int_a^t \epsilon(u)\frac{du}{u}\right)$$

has a "smoothly slowly varying" asymptotically equivalent version

$$\tilde{\ell}(t) = c \cdot \exp\left(\int_a^t \tilde{\epsilon}(u)\frac{du}{u}\right)$$

That is, $\ell(t)/\tilde{\ell}(t) \to 1$ *as* $t \to \infty$.

One of the main uses of regular variation is to obtain asymptotically equivalent versions of integrals. The main device is Karamata's Theorem, as stated in pg. 8 by [32]. See also pg. 26 of [7].

Theorem 1.1.10 (Karamata's Theorem)**.** *Let* R *be regularly varying of order* α *and locally bounded on* $[T, \infty)$*. Then*

$$\lim_{t\to\infty} \frac{\int_T^t \theta R(\theta)d\theta}{t^2 R(t)} = \frac{1}{|2+\alpha|}$$

Karamata's theorem is often used in tandem with H. Potter's bounds:

Theorem 1.1.11 (Potter's Bounds)**.** *Let* $\ell(t)$ *be slowly varying as* $t \to \infty$*. Then, given any* $\epsilon_* > 0, C > 1$*, there is a cutoff* t_0 *such that for* $t_0 \le t_1 \le t_2$*,*

$$C^{-1}(t_1/t_2)^{-\epsilon_*} \le \ell(t_1)/\ell(t_2) \le C(t_1/t_2)^{\epsilon_*} \tag{1.3}$$

1.2 Sparsity and Regularization

In the best case, $\xi = 0$ and rank $X = p$ (full column rank), so that the response is exact and the design X is left-invertible. Suppose, for example, that we can *accurately* sample a signal $f \in L_2[-\pi, \pi]$ and we want to approximate f by an element in $V = \text{span}\{\sin(kx), \cos(kx) : k = 1, 2\}$. Conventional wisdom tells us to sample f at least twice as many times as the highest frequency (see Nyquist rate [30, 31]). Recent developments in compressed sensing, however, acheive sub-Nyquist

rates when a secret weapon called sparsity is available. We sample at four different times $t_1, \ldots, t_4$, then solve the linear system

$$\begin{bmatrix} f(t_1) \\ \vdots \\ f(t_4) \end{bmatrix} = \begin{bmatrix} \sin(t_1) & \cos(t_1) & \sin(t_2) & \cos(2t_1) \\ \vdots & \vdots & \vdots & \vdots \\ \sin(t_4) & \cos(t_4) & \sin(2t_4) & \cos(2t_4) \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_4 \end{bmatrix}$$

or $Y = X\beta$. Except for select choices of $t_1 \ldots t_4$, the design X will have full column rank and we can solve β exactly. Then, β will represent the coordinates, with respect to the basis $\{\sin(kx), \cos(kx) : k = 1, 2\}$, for the unique element $\widetilde{f} \in V$ that matches f at our four points $\widetilde{f}(t_i) = f(t_i)$, for $1 \leq i \leq 4$. This solution is exact ($\sin(kx), \cos(kx)$ are orthogonal) so there is no need to talk about a best approximation.

Now, suppose we still accurately ($\xi = 0$) sample $f(t)$ at four times, but enlarge V to span$\{\sin(kx), \cos(kx) : k = 1, 2, 3\}$. The same approach leads to an underdetermined linear system with a 4×6 design matrix and infinitely many solutions that exactly fit the data. Occam's razor suggests we look for the "simplest" solution. To this end, we devise what is called a "regularized" estimator to enforce the kind of simplicity that we suspect is appropriate.

For $f \in L_2$, this could mean smallest norm, i.e. least total energy. The least energy solution would then be found through $\arg\min_{Y=Xu} \|u\|_2$. Another interpretation of simplicity is to have the fewest number of nonzero coordinates. This idea of simplicity is called "sparsity" and is appropriate in many modern and diverse contexts, including compressed sensing, computational biology, health care, and advertising. Formally, a vector u is said to have *sparsity* s (to be s-sparse) if it has at

most s nonzero coordinates. Then, the simplest solution becomes

$$\arg\min_{Y=Xu} \|u\|_0$$

where $\|u\|_0 = \lim_{q\to 0} \|u\|_q = \#\{j : u_j \neq 0\}$. Unfortunately, solving this minimization problem reduces to searching for a solution to $Y = Xu$ for every possible support of u, as $\|u\|_0$ increases. The complexity boosts from polynomial time to NP-Hard [20, 23], so it is imperative that we take another approach.

The main asset of $\|u\|_0$ comes from an embarrasing truth - it is not even homogeneous: $\|au\|_0 = \|u\|_0$ for $a \neq 0$. We want to replace $\|\cdot\|_0$ with a bona fide norm and keep the misbehavior, but improve computatational efficiency. The popular fix (justified in [9]) uses $\|\cdot\|_1$ as a proxy for $\|\cdot\|_0$, since $\|\cdot\|_1$ is the only q-norm enjoying both convexity and nondifferentiability at 0 (properties of $\|\cdot\|_2$ and $\|\cdot\|_0$, respectively). By regularizing with a convex norm, we are able to use efficient methods from convex programming (use [5] as reference). Nondifferentiability at 0 potentially causes small (at the noise level) estimates to shrink all the way to 0. This "ℓ^1-regularized" version is called basis pursuit:

$$\widetilde{\beta}_{\mathrm{BP}} = \arg\min_{Y=Xu} \|u\|_1$$

(Donoho and Huo, 2001; Feuer and Nemirovski, 2003). The analysis of basis pursuit is quite elegant. If β is s-sparse, i.e. $S = \{j : \beta_j \neq 0\}$ has cardinality s, the necessary and sufficient condition for basis pursuit to enjoy exact recovery $\widetilde{\beta}_{\mathrm{BP}} = \beta$ is called the restricted nullspace condition of order s (the terminology "restricted nullspace" is from [11], but the condition itself originates from [12, 16]):

$$Xu = 0 \Rightarrow \|u_{S^c}\|_1 > \|u_S\|_1$$

For our purposes, we will require the stronger "compatability" condition (first from [34] section 2.1, but [8] section 2 uses notation closer to ours) to allow noise. It replaces "$Xu = 0$" with "$\|Xu\|_2 < \zeta\|u\|_1$." The name compatability refers to the comparison between $\|\cdot\|_2$ and $\|\cdot\|_1$.

We will be particularly interested in sparsity as it pertains to solving linear systems, both when X has full column rank and when X is column rank deficient. We will also address noisy data, both in the response Y (standard linear regression) and in the design X (errors-in-variable regression, see [17]). Either way, when noise is present, exact solutions are no longer such a priority or even a possibility.

1.3 Noisy Problem

Ordinary least squares, or a carefully chosen variant thereof, often works well when the columns, X_j, of X are far from linearly dependent. Otherwise, a minute change in Y can cause a drastic change in the $\widetilde{\beta}_j$'s corresponding to the X_j's that are related.

To cope with noise, we no longer require exact solutions, but would ideally still like to enforce sparsity via the $\|\cdot\|_0$-norm. Of course, we don't know a priori the sparsity of β, so we have to guess the correct sparsity, typically with cross-validation (see [3]). Once a sparsity s is chosen, the corresponding subset selection estimator is defined as

$$\hat{\beta}_{\text{SS}} := \underset{\|u\|_0 \leq s}{\arg\min} \|Y - Xu\|^2$$

Not surprisingly, this problem is again NP-Hard [37], and we again use $\|\cdot\|_1$ as a proxy for $\|\cdot\|_0$. The ℓ_1-regularized regression estimator, which is usually known as the Least Absolute Shrinkage and Selection Operator (LASSO), or just basis-pursuit

denoising, is defined by

$$\hat{\beta}_n :\in \underset{\|u\|_1 \leq s}{\arg\min} \|Y - Xu\|^2$$

although the following dual problem is easier to compute: let λ be a constant (like s, to be determined by cross-validation) then

$$\hat{\beta}_n \in \underset{u \in \mathbb{R}^p}{\arg\min} \left[\|Xu - y\|_2^2 + \lambda \|u\|_1 \right] \tag{1.4}$$

The symbol $\in$ suggests there could be multiple solutions. In this work, which focuses on point estimation, the choice is immaterial. We care more about proximity than selecting the correct model. Therefore, when there are multiple solutions for $\hat{\beta}_n$, we won't prefer one over another, as long as they are close. At least, it turns out that $X\hat{\beta}$ is always unique, and so is $\|\hat{\beta}\|_1$ for $\lambda > 0$. For the problem of uniqueness and degrees of freedom see [33].

1.4 Equivalence of Limits of Sums

It has been said that mathematics never truly accomplishes anything. It merely combines trivialities, such as multiplying by 1 or adding 0. Indeed, our paper is a testament to this indictment. It is the crux of our next two lemmas, and our cornerstone. We use them to assert equivalence of limits of sums.

The first is a well known generalization of the fact

$$\frac{a}{b} = \frac{c}{d} \qquad \Longrightarrow \qquad \frac{a}{b} = \frac{c}{d} \;=\; \frac{a+c}{b+d}$$

Yet, we state it and give a proof for completeness sake. Consequently, a multiplication by 1 shows the asymptotic equivalence of sums of "proportionate" infinitesimals.

Lemma 1.4.1 (Times One Lemma)**.** *Let a_t, b_t be nonnegative functions of $t \in (0, \infty)$*

such that $a_t/b_t \to 1$ *as* $t \to \infty$ *and* $t = (t_{ni})_{i \leq n}$ *be a triangular array satisfying* $\min_{i \leq n} t_{ni} \to \infty$ *as* $n \to \infty$*, then*

$$\left(\sum_{i=1}^{n} a_{t_{ni}} \right) \Big/ \left(\sum_{i=1}^{n} b_{t_{ni}} \right) \to 1$$

It follows that if $\sum_{i=1}^{n} b_{t_{ni}}$ *converges, then* $\sum_{i=1}^{n} a_{t_{ni}}$ *converges to the same limit.*

Proof. Let $a, b \in [0, \infty)^n$. Since

$$\left| \sum_{i=1}^{n} (a_i - b_i) \right| = \left| \sum_{i=1}^{n} \left(\frac{a_i}{b_i} - 1 \right) b_i \right| \leq \left(\max_{i \leq n} \left| \frac{a_i}{b_i} - 1 \right| \right) \left(\sum_{i=1}^{n} b_i \right)$$

we get

$$\begin{aligned} \left| \left(\sum_{i=1}^{n} a_i \right) \Big/ \left(\sum_{i=1}^{n} b_i \right) - 1 \right| &= \left| \left(\sum_{i=1}^{n} (a_i - b_i) \right) \Big/ \left(\sum_{i=1}^{n} b_i \right) \right| \\ &\leq \max_{i \leq n} \left| \frac{a_i}{b_i} - 1 \right| \end{aligned}$$

That is, the percent error of a sum of approximations is no worse than the worst individual percent error. Thus,

$$\begin{aligned} \left| \left(\sum_{i=1}^{n} a_{t_{ni}} \right) \Big/ \left(\sum_{i=1}^{n} b_{t_{ni}} \right) - 1 \right| &\leq \max_{i \leq n} \left| \frac{a_{t_{ni}}}{b_{t_{ni}}} - 1 \right| \\ &\leq \sup_{r \geq \min t_{ni}} \left| \frac{a_r}{b_r} - 1 \right| \\ &\to 0 \end{aligned}$$

The last convergence follows from $a_t/b_t \to 1$. □

The second lemma aids in dealing with sums of regularly varying functions. Essentially, the total difference between a regularly varying function and the same

order power becomes negligible as the values get small. Consequently, an addition of 0 shows the asymptotic equivalence of sums of regularly varying infinitesimals.

Lemma 1.4.2 (Plus Zero Lemma)**.** *Suppose $\mathcal{R}^{-\alpha}(t)$ is regularly varying of order $-\alpha$ for some $\alpha \neq 0$, that b_n is a sequence of numbers so that $\mathcal{R}^{-\alpha}(b_n) = 1/n$. Let $W_-(t) = t^{\alpha-\kappa}\mathcal{R}^{-\alpha}$ and $W_+(t) = t^{\alpha+\kappa}\mathcal{R}^{-\alpha}$, regularly varying functions of orders $-\kappa$ and κ, respectively, for some $0 < \kappa < \alpha$. If a_i is a sequence of positive numbers so that*

$$\sup_n \frac{1}{n}\sum_{i=1}^{n} a_i^{\alpha+\kappa} < \infty,$$

then

$$\sum_{i=1}^{n}\left|\mathcal{R}^{-\alpha}(b_n/a_i) - \frac{a_i^\alpha}{n}\right| \to 0 \qquad \text{as } n \to \infty$$

Proof. By definition of b_n,

$$\begin{aligned}
\sum_{\substack{i\leq n\\ 1\leq a_i}}\left|\mathcal{R}^{-\alpha}(b_n/a_i) - \frac{a_i^\alpha}{n}\right| =& \frac{1}{n}\sum_{\substack{i\leq n\\ 1\leq a_i}}\left|\frac{\mathcal{R}^{-\alpha}(b_n/a_i)}{\mathcal{R}^{-\alpha}(b_n)} - a_i^\alpha\right| \\
=& \frac{1}{n}\sum_{\substack{i\leq n\\ 1\leq a_i}} a_i^{\alpha+\kappa}\left|\frac{W_+(b_n/a_i)}{W_+(b_n)} - a_i^{-\kappa}\right| \\
\leq& \left(\sup_n \frac{1}{n}\sum_{\substack{i\leq n\\ 1\leq a_i}} a_i^{\alpha+\kappa}\right)\sup_{1\leq t}\left|\frac{W_+(b_n/t)}{W_+(b_n)} - t^{-\kappa}\right| \\
\to& 0
\end{aligned}$$

by the Uniform Convergence Theorem 1.1.5. Likewise,

$$\begin{aligned}
\sum_{\substack{i\le n\\ a_i\le 1}} \left|\mathcal{R}^{-\alpha}(b_n/a_i) - \frac{a_i^\alpha}{n}\right| =& \frac{1}{n}\sum_{\substack{i\le n\\ a_i\le 1}} \left|\frac{\mathcal{R}^{-\alpha}(b_n/a_i)}{\mathcal{R}^{-\alpha}(b_n)} - a_i^\alpha\right| \\
=& \frac{1}{n}\sum_{\substack{i\le n\\ a_i\le 1}} a_i^{\alpha-\kappa}\left|\frac{W_-(b_n/a_i)}{W_-(b_n)} - a_i^\kappa\right| \\
\le& \left(\sup_n \frac{1}{n}\sum_{\substack{i\le n\\ a_i\le 1}} a_i^{\alpha-\kappa}\right)\sup_{t\le 1}\left|\frac{W_-(b_n/t)}{W_-(b_n)} - t^\kappa\right| \\
\to& 0
\end{aligned}$$

□

1.5 Centering and Scaling the LASSO

Assume $(\xi_i)_{i\ge 1}$ are i.i.d. and that both

$$R^{-\alpha}(t) := \mathbb{P}\{\xi_i > t\} \quad \text{and} \quad r^{-\alpha}(t) := \mathbb{P}\{-\xi_i > -t\} \tag{1.5a}$$

are regularly varying of order $-\alpha$ as $t \to \infty$. With no loss of generality (see Remark 1.1.9), assume $R^{-\alpha}$ and $r^{-\alpha}$ are continuous. Define

$$b_n := \inf\{t \ge 0 : R^{-\alpha}(t) + r^{-\alpha}(t) = 1/n\} \tag{1.5b}$$

We want to work with a centered and scaled version of $\hat{\beta}_n$. To this end, substitute $Y = X\beta + \xi$ into the definition of $\hat{\beta}_n$, equation (1.4)

$$\begin{aligned}\hat{\beta}_n &:= \arg\min_{u\in\mathbb{R}^p} \left[\|Xu - Y\|^2 + \lambda_n\|u\|_1\right] \\ &= \arg\min_{u\in\mathbb{R}^p} \left[\|Xu - (X\beta + \xi)\|^2 + \lambda_n\|u\|_1\right] \\ &= \arg\min_{u\in\mathbb{R}^p} \left[\|X(u-\beta) - \xi\|^2 + \lambda_n\|u\|_1\right]\end{aligned}$$

Then, write the above in terms of a local parameter. That is, u becomes $u + \beta$.

$$\hat{\beta}_n = \beta + \arg\min_{u\in\mathbb{R}^p} \left[\|Xu - \xi\|^2 + \lambda_n\|u + \beta\|_1\right]$$

Expand the square and subtract $\|\xi\|^2 + \lambda_n\|\beta\|_1$, which does not depend on u, hence does not affect the arg min.

$$\hat{\beta}_n - \beta = \arg\min_{u\in\mathbb{R}^p} \left[\|Xu\|^2 - 2\langle Xu, \xi\rangle + \lambda_n(\|u + \beta\|_1 - \|\beta\|_1)\right]$$

Scale the parameter by b_n^{-1} (which is to say, u becomes $b_n^{-1}u$).

$$\begin{aligned}\hat{\beta}_n - \beta = b_n^{-1} \arg\min_{u\in\mathbb{R}^p} \big[&b_n^{-2}\|Xu\|^2 - 2b_n^{-1}\langle Xu, \xi\rangle \\ &+\lambda_n(\|b_n^{-1}u + \beta\|_1 - \|\beta\|_1)\big]\end{aligned}$$

Finally, we arrive at the relevant random quantity

$$\begin{aligned}\hat{u}_n := b_n(\hat{\beta}_n - \beta) = \arg\min_{u\in\mathbb{R}^p} \big[\ &b_n^{-2}\|Xu\|^2 - 2b_n^{-1}\langle u, X^T\xi\rangle \\ &+\lambda_n b_n^{-1}(\|u + b_n\beta\|_1 - \|b_n\beta\|_1)\big]\end{aligned} \tag{1.6}$$

Now, look at the ℓ^1 part. Firstly, if $\beta_j = 0$, then

$$|u_j + b_n\beta_j| - |b_n\beta_j| = |u_j|$$

Secondly, if $\text{Sign}(u_j) = -\text{Sign}(\beta_j) \neq 0$ and $|u_j| > b_n|\beta_j|$ (equivalently $b_n\beta_j$ is between $-u_j$ and 0), then

$$\begin{aligned}
|u_j + b_n\beta_j| - |b_n\beta_j| &= (u_j + b_n\beta_j)\text{Sign}(u_j) - (b_n\beta_j)\text{Sign}(\beta_j) \\
&= u_j\text{Sign}(-u_j) + 2u_j\text{Sign}(u_j) + b_n\beta_j(\text{Sign}(u_j) - \text{Sign}(\beta_j)) \\
&= u_j\text{Sign}(\beta_j) + 2|u_j| - 2b_n|\beta_j|
\end{aligned}$$

Thirdly, if either $\text{Sign}(u_j) = \text{Sign}(\beta_j) \neq 0$ or $|u_j| \leq b_n|\beta_j|$,

$$\begin{aligned}
|u_j + b_n\beta_j| - |b_n\beta_j| &= (u_j + b_n\beta_j)\text{Sign}(\beta_j) - (b_n\beta_j)\text{Sign}(\beta_j) \\
&= u_j\text{Sign}(\beta_j)
\end{aligned}$$

So, using as a shorthand for the Gram matrix, cross term, and the so-called "catastrophic correction" term

$$C_n := b_n^{-2} X_n^T X_n \tag{1.7a}$$

$$Z_n := b_n^{-1} \sum_{i=1}^{n} x_i \xi_i \tag{1.7b}$$

$$\mathcal{E}_n(u) := 2 \sum_{j=1}^{p} (|u_j| - b_n|\beta_j|)_+ \mathbb{1}\{\text{Sign}(u_j\beta_j) = -1\} \tag{1.7c}$$

with $C_n \in \mathbb{R}^{p\times p}$, $Z_n \in \mathcal{L}_1(\mathbb{R}^p)$, and $\mathcal{E}_n(u) : \mathbb{R}^p \to \mathbb{R}$. Equation (1.6) becomes

$$\hat{u}_n = \arg\min V_n(u)$$

where

$$V_n(u) = \langle u, C_n u\rangle - 2\langle u, Z_n\rangle + \lambda_n/b_n \sum_{j=1}^{p} \left\{ \begin{array}{ll} u_j \text{Sign}(\beta_j) & \text{if} \beta_j \neq 0 \\ |u_j| & \text{if} \beta_j = 0 \end{array} \right\} + \lambda_n/b_n \mathcal{E}_n(u) \tag{1.8}$$

Note: We mostly work within one row of a triangular array at a time. Dependence on n tends to be suppressed throughout this work, coming back out only when we pass to limits.

Simply note that the function inside the arg min above evaluates to 0 when $u = 0$, and that $\mathcal{E}_n(u) \geq 0$ for all $u \in \mathbb{R}^p$. This begets the Basic Inequality,

$$\langle \hat{u}, C\hat{u}\rangle - 2\langle \hat{u}, Z_n\rangle + \lambda_n b_n \sum_{j=1}^{p} \left\{ \begin{array}{ll} \hat{u}_j \text{Sign}(\beta_j) & \text{if } \beta_j \neq 0 \\ |\hat{u}_j| & \text{if } \beta_j = 0 \end{array} \right\} \leq 0 \tag{1.9-BI}$$

To ensure that changing a single covariate is asymptotically inconsequential, we take $\|x_n\|_\infty = o(b_n)$. Equivalently,

$$b_n^{-1} \sup_{i \leq n} \|x_i\| \to 0 \tag{1.10}$$

so that the summands $b_n^{-1}\xi_i x_i$ are uniformly infinitesimal:

$$\lim_{n\to\infty} \sup_{i \leq n} \mathbb{P}\{\|b_n^{-1}\xi_i x_i\|_\infty > \epsilon\} = 0$$

for every $\epsilon > 0$. Also, to keep the sums nearly centered, assume either

$$\mathbb{E}\xi_i = 0 \qquad \text{or} \qquad b_n^{-1} \sum_{i=1}^{n} x_i \to 0 \tag{1.11}$$

with $t_{ni} := b_n / \|x_i\|$.

1.6 The Argmin Theorem

To pass from convergence in law of the LASSO objective to convergence in law of the argmin, we need a version of the continuous mapping theorem (CMT). That is, if $\arg\min : \ell^\infty(U) \to U$ were continuous, we could apply CMT. Though, requiring continuity of $\arg\min$, even just locally, is often too strong. Theorem 3.2.2 in [35] requires instead uniform tightness of the sequence of argmins, which is clearly necessary. We state a reduced form, then distill the proof from [35].

Theorem 1.6.1. *Let $\mathbb{M}_n, \mathbb{M}$ be stochastic processes indexed by a metric space U such that $\mathbb{M}_n \rightsquigarrow \mathbb{M}$ in $\ell^\infty(K)$ for every compact $K \subset U$. Suppose that almost all sample paths $u \to \mathbb{M}(u)$ are lower semicontinuous and possess a unique minimum at a random point $\hat{u}_\infty$, which as a random map in U is tight. If the sequence $\hat{u}_n \in \arg\min_{u \in U} \mathbb{M}_n(u)$ is uniformly tight, then $\hat{u}_n \rightsquigarrow \hat{u}$ in U (the symbol $\rightsquigarrow$ denotes weak convergence).*

Proof. The portmanteau theorem states that $\hat{u}_n \rightsquigarrow \hat{u}_\infty$ is equivalent to

$$\overline{\lim}\, \mathbb{P}\{\hat{u}_n \in F\} \leq \mathbb{P}\{\hat{u}_\infty \in F\}$$

for every closed $F \subset U$. By hypothesis, for every $\epsilon > 0$, we have a compact $K \subset U$

such that $\mathbb{P}\{\hat{u}_n \notin K \text{ or } \hat{u}_\infty \notin K\} < \epsilon$. Hence

$$\begin{aligned}\mathbb{P}\{\hat{u}_n \in F\} \leq & \mathbb{P}\{\hat{u}_n \in F \cap K\} + \epsilon \\ \leq & \mathbb{P}\left\{\inf_{F\cap K} \mathbb{M}_n(h) \leq \inf_K \mathbb{M}_n(h)\right\} + \epsilon \end{aligned} \tag{1.12}$$

$\inf_{F\cap K}(\cdot)$ and $\inf_K(\cdot)$ are continuous mappings $\ell^\infty(U) \to \mathbb{R}$. By the continuous mapping theorem,

$$\inf_K(\mathbb{M}_n(u)) - \inf_{F\cap K}(\mathbb{M}_n(u)) \rightsquigarrow \inf_K(\mathbb{M}(u)) - \inf_{F\cap K}(\mathbb{M}(u))$$

which by the portmanteau theorem (and closedness of $[0, \infty)$) means

$$\begin{aligned}\overline{\lim}\, \mathbb{P}\left\{\inf_{F\cap K} \mathbb{M}_n(h) \leq \inf_K \mathbb{M}_n(h)\right\} \leq & \mathbb{P}\left\{\inf_{F\cap K} \mathbb{M}(h) \leq \inf_K \mathbb{M}(h)\right\} \\ = & \mathbb{P}\{\hat{u}_\infty \in F \cap K\} \\ \leq & \mathbb{P}\{\hat{u}_\infty \in F\} + \epsilon \end{aligned} \tag{1.13}$$

Taking $\overline{\lim}$ across (1.12) and pairing with (1.13), we get

$$\overline{\lim}\, \mathbb{P}\{\hat{u}_n \in F\} \leq \mathbb{P}\{\hat{u}_\infty \in F\} + 2\epsilon$$

But ϵ only depended on K. □

2. FIXED DESIGN, FIXED NUMBER OF REGRESSORS

We start our analysis of the LASSO with a single design X and a fixed number of regressors X_j (columns of X). For consistency results, we mention some laws of large numbers.

2.1 Laws of Large Numbers

The classical statement of the weak law of large numbers (WWLN) is stated in terms of moments, but moments are not necessary. Interestingly, Feller believed (in [14] p.152) the WLLN to be "of limited interest and should be replaced by the more precise and more useful strong law of large numbers," contrary to van der Waerden's later statement, "[The strong law of large numbers] scarcely plays a role in mathematical statistics." ([36] p.98). Even so, Feller gave necessary and sufficient conditions for the WLLN in the i.i.d. case (see [15]),

Theorem 2.1.1 (Feller's WLLN)**.** *Let ξ_i be i.i.d. In order that there exist constants a_n so that for every $\epsilon > 0$,*

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i - a_n\right| > \epsilon\right\} \to 0$$

it is necessary and sufficient that

$$t\mathbb{P}\{|\xi_i| > t\} \to 0 \tag{2.1}$$

as $t \to \infty$. In this case, it suffices to set $a_n = \mathbb{E}[\ |\xi_i|\mathbb{1}\{|\xi_i| < n\}\ ]$

Of course, the classical hypothesis for the WLLN, $\mathbb{E}|\xi_i| < \infty$, implies

$$t\mathbb{P}\{|\xi_i| > t\} \quad \leq \quad \mathbb{E}[\ |\xi_i|\mathbb{1}\{|\xi_i| > t\}\ ] \quad \to \quad 0$$

as $t \to \infty$. The converse is not true in general, which is seen by considering $\mathbb{P}\{|\xi_i| > t\} = (t \log t)^{-1}$ for $t \geq 2$:

$$\mathbb{E}|\xi_i| = \int_2^\infty (t \log t)^{-1} dt = \log(\log t)\Big|_2^\infty = \infty$$

Just as with the weighted (nonidentical) Marcinkiewicz SLLN in [10], we would like a weighted version of Feller's WLLN. With this in mind, we can break (2.1) into an equivalent sum-of-infinitesimals version to reflect the "average" tail behavior when ξ_i are not identical. That is, if $n - 1 \leq t \leq n$, then

$$\frac{n-1}{n} \cdot n\mathbb{P}\{|\xi_i| > n\} \quad \leq \quad t\mathbb{P}\{|\xi_i| > t\} \quad \leq \quad \frac{n}{n-1} \cdot (n-1)\mathbb{P}\{|\xi_i| > n-1\}$$

So, (2.1) is equivalent to

$$n\mathbb{P}\{|\xi_i| > n\} = \sum_{i=1}^{n} \mathbb{P}\{|\xi_i| > n\} \to 0$$

as $n \to \infty$. From Feller's WLLN, we could guess the weighted WLLN, which we present as a corollary to Theorem 2.2.7 (next subsection). Then, as an appetizer, we prove consistency of OLS for certain distributions of ξ_i where $\mathbb{E}|\xi_i| = \infty$.

Corollary 2.1.2 (Weighted WLLN)**.** *Let ξ_i be i.i.d. and $x_i \in \mathbb{R}^p$ for $i \in \mathbb{N}$ so that*

$$\frac{1}{n} \max_{i \leq n} \|x_i\| \to 0 \qquad \text{and} \qquad \sup_n \frac{1}{n} \sum_{i=1}^{n} \|x_i\| < \infty \tag{2.2}$$

For there to exist constants a_{ni} so that for every $\epsilon > 0$,

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}(x_i\xi_i - a_{ni})\right| > \epsilon\right\} \to 0$$

it is necessary and sufficient that (2.1) hold. In this case, it suffices to set

$$a_{ni} = x_i \cdot \mathbb{E}[\ |\xi_i|\mathbb{1}\{|\xi_i| < n/\|x_i\|\}\ ]$$

Proof. Write B_r for the ball of radius r centered at the origin. Setting $Z_{ni} = x_i\xi_i/n$ in Theorem 2.2.7 (next subsection), we check (2.5a) by computing $\Phi(\mathbb{R}^p \setminus B_r)$.

$$\begin{aligned}
\Phi(\mathbb{R}^p \setminus B_r) &= \lim_{n\to\infty}\sum_{i=1}^{n}\mathbb{P}\{\|Z_{ni}\| > r) \\
&= \lim_{n\to\infty}\sum_{i=1}^{n}\mathbb{P}\left\{|\xi_i| > \frac{nr}{\|x_i\|}\right\} \\
&= \lim_{n\to\infty}\frac{1}{nr}\sum_{i=1}^{n}\|x_i\|\left(\frac{nr}{\|x_i\|}\mathbb{P}\left\{|\xi_i| > \frac{nr}{\|x_i\|}\right\}\right) \\
&= 0
\end{aligned}$$

which comes after considering (2.1), $\min_{i\le n}(nr/\|x_i\|) \to \infty$, and $\sup_n \frac{1}{n}\sum_{i=1}^{n}\|x_i\| < \infty$. It follows that $\Phi(E) = 0$ for every Borel set $E \subset \mathbb{R}^p \setminus \{0\}$. For the Gaussian component $Q(u)$ in (2.5a), normalize by setting $\|u\| = 1$. Then, by the Cauchy-Schwarz inequality,

$$\begin{aligned}
\mathrm{Var}(\langle Z_{ni}\mathbb{1}\{\|Z_{ni}\| \le r\}, u\rangle) &\le \frac{\|x_i\|^2}{n^2}\mathrm{Var}\left(|\xi_i|\mathbb{1}\left\{|\xi_i| \le \frac{nr}{\|x_i\|}\right\}\right) \\
&\le \frac{\|x_i\|^2}{n^2}\int_0^{nr/\|x_i\|}\mathbb{P}\{|\xi_i| > t\}\cdot 2t dt
\end{aligned}$$

then using the facts

$$\lim_{w\to\infty} \frac{1}{w}\int_0^w f(t)dt \to 0 \qquad \text{if} \qquad \lim_{t\to\infty} f(t) = 0$$

and

$$\frac{1}{n}\sum_{i=1}^{n} a_i \to 0 \qquad \text{if} \qquad \sup_{i\le n} |a_i| \to 0$$

we have

$$2r \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{\|x_i\|}{nr} \int_0^{nr/\|x_i\|} t\mathbb{P}\{|\xi_i| > t\}dt \to 0$$

showing $Q(u) = 0$ for every $u \in \mathbb{R}^p$. □

Theorem 2.1.3. *Suppose ξ_i are i.i.d., satisfying (2.1). If $\frac{1}{n}X_n^T X_n =: C_n \to C_\infty$ in norm for some positive definite $C_\infty \in \mathbb{R}^{p\times p}$. Then, the OLS estimator, $\hat{\beta}_{n,\mathrm{OLS}} = (X_n^T X_n)^{-1} X_n^T Y$ is consistent. That is, there are nonrandom centerings a_n such that*

$$\hat{\beta}_{n,\mathrm{OLS}} - a_n \xrightarrow{\mathbb{P}} \beta$$

in norm.

Proof. Suppress the dependence of X on n and set

$$a_n = \left(X^T X\right)^{-1} x_i \cdot \mathbb{E}[\ |\xi_i| \mathbb{1}\{|\xi_i| < n/\|x_i\|\}\ ]$$

Substitute $Y = X\beta + \xi$ into the normal equation to get

$$\begin{aligned}\hat{\beta}_{n,\text{OLS}} - \beta - a_n =& (X^T X)^{-1} X^T X\beta + (X^T X)^{-1} X^T \xi - a_n - \beta \\ =& (X^T X)^{-1} \left(X^T \xi - x_i \cdot \mathbb{E}[\ |\xi_i| \mathbb{1}\{|\xi_i| < n/\|x_i\|\}\ ] \right) \\ =& \left(\frac{1}{n} X^T X \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x_i \xi_i - a_{ni} \right) \end{aligned} \tag{2.3}$$

with a_{ni} as in Corollary 2.1.2. Now, let $\eta > 0$ be the smallest eigenvalue of C_∞. Note, convergence of C_n in norm implies convergence of ordered eigenvalues. So, the smallest eigenvalue of C_n must converge to η and eventually be greater than $\eta/2$. By Corollary 2.1.2, it follows that for large n,

$$\left\| \left(\frac{1}{n} X^T X \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x_i \xi_i - a_{ni} \right) \right\| \leq \frac{2}{\eta} \left\| \left(\frac{1}{n} \sum_{i=1}^{n} x_i \xi_i - a_{ni} \right) \right\|$$

and the right side goes to 0 in probability. Referring back to equation (2.3), the proof is finished. □

2.2 Infinitely Divisible Central Limit Theorem in $\mathbb{R}^p$

In 1837, Siméon Poisson published a book [26] that included a study of the number of wrongful convictions in a country over a certain period. Suppose on any day there is a probability 1/365 that a certain country experiences a wrongful conviction, that the number of convictions on different days are independent of each other, and that it is highly unlikely to have two or more in one day. The total number of convictions in one year can then be modeled as the sum of 365 Bernoulli variables.

$$\begin{aligned}\text{Binom}\left(1, \frac{1}{365}\right) + \cdots + \text{Binom}\left(1, \frac{1}{365}\right) \sim& \text{Binom}\left(365, \frac{1}{365}\right) \\ \approx& \text{Pois}\left(365 \cdot \frac{1}{365}\right)\end{aligned}$$

This approximation is sometimes called the law of rare events, where 'law' is a synonym for distribution and convictions are considered 'rare events'.

The law of rare events shows how a Poisson distribution can be broken into an independent sum of many usually-zero independent random variables, a property called infinite divisibility. The idea of usually-zero random variables will become important when discussing the Lévy measure from Theorem 1.1.2 and compound Poisson processes.

Traditionally, the "stable CLT" is stated in terms of i.i.d. random variables. While we assume i.i.d. errors ξ_i, we also assume the covariates x_i are fixed (not i.i.d.). As with most nonparametric regressions, we are concerned with a normalized version of $\sum_{i=1}^{n} \xi_i x_i$ (not i.i.d.). Hence, we must work with a more general CLT involving non-i.i.d. terms. The multivariate version of the CLT in Ch. 25 of [18] for infinitely divisible triangular arrays was proved in [29]. Unfortunately, [29] presumed a multivariate Khintchine representation, which was incorrect.

We follow the univariate treatment in chapter 2 of [2] to give a new proof. We do not take any credit, for the proofs are the same. First, we present a method called the "decoupage de Lévy," which approximately decomposes a usually small random variable, Z, into an always small variable, $Z^\flat$, and a usually-zero variable, $Z^\#$. Generally, small parts tend to cumulatively behave like normal distributions, and, as will be shown (Lemma 2.2.6), usually-zero-parts tend to cumulatively behave like Poisson distributions.

Lemma 2.2.1. *Let Z be a random vector and E a Borel set in $\mathbb{R}^p$ bounded away from 0. Let $Z^\# = Z|_E$, $Z^\flat = Z|_{E^c}$, and W be independent of Z with Bernoulli distribution and mean $\mathbb{P}\{Z \in E\}$. Then,*

$$\mathcal{L}(Z) = \mathcal{L}\left(WZ^\# + (1-W)Z^\flat\right)$$

Proof. For any Borel F,

$$\begin{aligned}\mathbb{P}\{WZ^{\#} + (1-W)Z^{\flat} \in F\} =& \mathbb{P}\{WZ^{\#} + (1-W)Z^{\flat} \in E \cap F\} \\ &+ \mathbb{P}\{WZ^{\#} + (1-W)Z^{\flat} \in E^c \cap F\} \\ =& \mathbb{P}\{Z^{\#} \in E \cap F\}\mathbb{P}\{W = 1\} \\ &+ \mathbb{P}\{Z^{\flat} \in E^c \cap F\}\mathbb{P}\{W = 0\} \\ =& \mathbb{P}\{Z \in E \cap F\} + \mathbb{P}\{Z \in E^c \cap F\} \\ =& \mathbb{P}\{Z \in F\}\end{aligned}$$

□

Even though W is independent of Z in Lemma 2.2.1, $WZ^{\#}$ and $(1-W)Z^{\flat}$ are still dependent (disjoint supports). To achieve independence in the proof of Theorem 2.2.7, we will introduce a residual term whose size will depend on the size of E^c. First, let us give some definitions and lemmas.

Definition 2.2.2. *Suppose ν is a finite Borel measure on $\mathbb{R}^p$. Define the characteristic function of ν by*

$$(\text{ch.f.}(\nu))(t) := \int e^{i\langle t,x\rangle} d\nu(x)$$

Write $|\nu|$ for its total variation, ν^k for its k-fold convolution (set $\nu^0 = \delta_0$, the point mass at 0), and $\text{Pois}(\nu)$ *for the compound Poisson distribution associated with ν (in $\mathbb{R}^p$):*

$$\text{Pois}(\nu) := e^{-|\nu|} \sum_{k=0}^{\infty} \frac{\nu^k}{k!}$$

Remark 2.2.3. *Writing $\nu = |\nu| \cdot \nu/|\nu|$, we see that if Z_i are i.i.d. with distribution $\nu/|\nu|$, then the random sum $\sum_{i=1}^{N} Z_i$ (with $N \sim \text{Pois}(|\nu|)$ and independent from Z_i) has distribution* $\text{Pois}(\nu)$.

Proof.

$$\begin{aligned}\int_E d\text{Pois}(\nu) =& e^{-|\nu|}\sum_{k=0}^{\infty}\frac{1}{k!}\int_E d\nu^k \\ =& e^{-|\nu|}\sum_{k=0}^{\infty}\frac{|\nu|^k}{k!}\int_E d\left(\frac{\nu}{|\nu|}\right)^k \\ =& \mathbb{P}[Z_1+\cdots+Z_N\in E]\end{aligned}$$

□

Realizing $\text{Pois}(\nu)$ as a random sum is convenient both in computation and as a reference model. For example, in the single variable ($p=1$) case, if the number of cars passing by a booth is $\text{Pois}(|\nu|)$ distributed and the amounts of marijuana in each car are i.i.d. distributed as $\nu/|\nu|$, then the total amount of marijuana passing by the booth is distributed as $\text{Pois}(\nu)$.

Now, we state some basic facts about compound Poisson distributions.

Lemma 2.2.4. *If ν_n, ν are finite Borel measures on $\mathbb{R}^p$, then*

(i) The distribution $\text{Pois}(\nu)$ *is a probability measure and*

$$\text{ch.f.}(\text{Pois}(\nu)) = \exp(\text{ch.f.}(\nu) - |\nu|) \tag{2.4}$$

(ii) $\text{Pois}(\nu)$ *has mean* $\int x d\nu(x)$ *and covariance* $\int xx^T d\nu(x)$*. Hence, the square norm of a* $\text{Pois}(\nu)$ *RV has variance* $\int \|x\|^2 d\nu(x)$*.*

(iii) If $\nu_n \rightsquigarrow \nu$*, then* $\text{Pois}(\nu_n) \rightsquigarrow \text{Pois}(\nu)$*.*

(iv) $\text{Pois}\left(\sum_{i=1}^n \nu_i\right) = \text{Pois}(\nu_1) * \ldots * \text{Pois}(\nu_n)$

(v) $\sup_{A\in\mathcal{B}} |\text{Pois}(\nu(A)) - \nu(A)| \leq (\nu(\mathbb{R}^p \setminus \{0\}))^2$

Proof. Let $N \sim \text{Pois}(|\nu|)$ and $Z_i \sim \nu/|\nu|$, so that we have the random sum representation $\sum_{i=1}^{N} Z_i \sim \text{Pois}(\nu)$. Then,

(i)

$$\begin{aligned}
\text{ch.f.}(\text{Pois}(\nu)) =& \mathbb{E} \exp\left(i \left\langle t, \sum_{i=1}^{N} Z_i \right\rangle \right) \\
=& \mathbb{E}[(\text{ch.f.}(\nu/|\nu|))^N] \\
=& e^{-|\nu|} \sum_{k=0}^{\infty} \frac{|\nu|^k}{k!} \cdot (\text{ch.f.}(\nu/|\nu|))^k \\
=& \exp(\text{ch.f.}(\nu) - |\nu|)
\end{aligned}$$

(ii)

$$\mathbb{E}\left(\sum_{i=1}^{N} Z_i \right) = \mathbb{E}\left(\mathbb{E}\left(\sum_{i=1}^{N} Z_i \middle| N \right) \right) = \mathbb{E}(N)\mathbb{E}(Z_1) = |\nu| \int x d\left(\frac{\nu}{|\nu|} \right)(x)$$

Also, since $\mathbb{E}(N) = |\nu|$ and $\mathbb{E}(N^2 - N) = |\nu|^2$,

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{i=1}^{N} Z_i \right) \left(\sum_{i=1}^{N} Z_i \right)^T \right] =& \mathbb{E}\left[\mathbb{E}\left(\sum_{i=1}^{N} \sum_{l=1}^{N} Z_i Z_l^T \middle| N \right) \right] \\
=& \mathbb{E}\left[\mathbb{E}\left(\sum_{\substack{i=l \\ i \le N}} + \sum_{\substack{i \neq l \\ i,l \le N}} Z_i Z_l^T \middle| N \right) \right] \\
=& |\nu| \mathbb{E}(Z_i Z_i^T) + |\nu|^2 (\mathbb{E}(Z_i)\mathbb{E}(Z_i)^T) \\
=& |\nu| \mathbb{E}(Z_i Z_i^T) + \mathbb{E}\left(\sum_{i=1}^{N} Z_i \right) \mathbb{E}\left(\sum_{i=1}^{N} Z_i \right)^T
\end{aligned}$$

Taking the trace of both sides yields the next statement.

(iii) By Lévy's Continuity Theorem and Lemma 2.2.4(i),

$$\begin{aligned}\text{ch.f. Pois}(\nu_n) =& \exp((\text{ch.f. } \nu_n) - a) \\ \to& \exp((\text{ch.f. } \nu) - a) \\ =& \text{ch.f. Pois}(\nu)\end{aligned}$$

(iv) By Lemma 2.2.4(i),

$$\begin{aligned}\text{ch.f. Pois}\left(\sum_{i=1}^{n}\nu_i\right) =& \exp\left(\left(\text{ch.f. } \sum_{i=1}^{n}\nu_i\right) - \sum_{i=1}^{n}|\nu_i|\right) \\ =& \text{ch.f. Pois}(\nu_1)\cdots\text{ch.f. Pois}(\nu_n)\end{aligned}$$

(v) Set $\epsilon = \nu(\mathbb{R}^p \setminus \{0\})$ and $\nu_0 = \nu\{0\}\delta_0$ and $\nu_1 = \nu - \nu_0$. Since $\text{Pois}(a\delta_0) = \delta_0$, we have by part (ii) that $\text{Pois}(\nu) = \text{Pois}(\nu\{0\}\delta_0 + \nu_1) = \text{Pois}(\nu_1)$. Now, for a general Borel set A, first consider the case $\text{Pois}(\nu)(A) \geq \nu(A)$.

$$\begin{aligned}0 \leq \text{Pois}(\nu)(A) - \nu(A) =& e^{-\epsilon}\sum_{k=0}^{\infty}\frac{\nu^k(A)}{k!} - (1-\epsilon)\delta_0(A) - \nu(A) \\ \leq& (e^{-\epsilon} + \epsilon - 1)\delta_0(A) + \sum_{k=2}^{\infty}\frac{\epsilon^k}{k!} \\ \leq& \frac{\epsilon^2}{2} + \frac{\epsilon^2}{2}\end{aligned}$$

Secondly, consider the case $\nu(A) \geq \text{Pois}(\nu)(A)$

$$\begin{aligned}0 \leq \nu(A) - \text{Pois}(\nu)(A) =& (1-\epsilon)\delta_0(A) + \nu(A) - e^{-\epsilon}\sum_{k=0}^{\infty}\frac{\nu^k(A)}{k!} \\ \leq& (1 - e^{-\epsilon})\nu(A) \\ \leq& \epsilon^2\end{aligned}$$

□

The next lemma is similar in spirit to Lemma 1.4.1, bounding the approximation error of a sum by a function of the individual approximation errors. See [2] for the univariate case.

Lemma 2.2.5. *Let $\mathcal{F}$ be a family of real bounded Borel functions on $\mathbb{R}^p$ and μ_i, ν_i Borel probability measures on $\mathbb{R}^p$. Then if $\mathcal{F}$ is closed under translations,*

$$\sup_{f\in\mathcal{F}}\left|\int f d(\mu_1 * \cdots \mu_n - \nu_1 * \cdots * \nu_n)\right| \leq \sum_{i=1}^{n} \sup_{f\in\mathcal{F}}\left|\int f d(\mu_i - \nu_i)\right|$$

Proof. It suffices to prove the lemma when $n = 2$. By definition of the convolution of measures, $\int f d(\mu * \nu) = \int\int f(x+y)\mu\{dx\}\nu\{dy\}$. Then,

$$\begin{aligned}
\sup_{f\in\mathcal{F}}\left|\int f d(\mu_1 * \mu_2 - \nu_1 * \nu_2)\right| &= \sup_{f\in\mathcal{F}}\left|\int\int f(x+y)(\mu_1\{dx\}\mu_2\{dy\} - \nu_1\{dx\}\nu_2\{dy\})\right| \\
&\leq \sup_{f\in\mathcal{F}}\left|\int\int f(x+y)(\mu_1 - \nu_1)\{dx\}\mu_2\{dy\}\right| \\
&\quad + \sup_{f\in\mathcal{F}}\left|\int\int f(x+y)(\mu_2 - \nu_2)\{dx\}\nu_1\{dy\}\right| \\
&\leq \int \sup_{f\in\mathcal{F}}\left|\int f d(\mu_1 - \nu_1)\right| d\mu_2 + \int \sup_{f\in\mathcal{F}}\left|\int f d(\mu_2 - \nu_2)\right| d\nu_1 \\
&\leq \sup_{f\in\mathcal{F}}\left|\int f d(\mu_1 - \nu_1)\right| + \sup_{f\in\mathcal{F}}\left|\int f d(\mu_2 - \nu_2)\right|
\end{aligned}$$

□

Theorem 2.2.6. *For any $n \in \mathbb{N}$, let $\{Z_i\}_{i\leq n}$ be independent $\mathbb{R}^p$-valued random variables. Then,*

$$\sup_{A\in\mathcal{B}}\left|\mathcal{L}\left(\sum_{i=1}^{n} Z_i\right)(A) - \text{Pois}\left(\sum_{i=1}^{n}\mathcal{L}(Z_i)\right)(A)\right| \leq \sum_{i=1}^{n}(\mathbb{P}\{Z_i \neq 0\})^2$$

Proof. Let $\mathcal{F}$ be the set of indicator functions of Borel sets, then apply Lemmas 2.2.4(iv,v) and 2.2.5 □

Now, we are ready to prove the infinitely divisible CLT for $\mathbb{R}^p$-valued random variables.

Theorem 2.2.7. *Let $(Z_{ni})_{i\leq n}$ be a uniformly infinitesimal triangular array of random vectors in $\mathbb{R}^p$ with independent rows. Define the truncation $Z_{ni\delta} := Z_{ni}\mathbb{1}\{\|Z_{ni}\| \leq \delta\}$ and $Z_{ni}^{\delta} := Z_{ni}\mathbb{1}\{\|Z_{ni}\| > \delta\}$. For there to exist $a_{ni}, a \in \mathbb{R}^p$ so that the sums $\sum_{i=1}^{n}(Z_{ni} - a_{ni})$ converge to an infinitely divisible distribution with Lèvy representation $[a, Q, \Phi]$, it is necessary and sufficient that*

(i) Φ is a Borel measure on $\mathbb{R}^p \setminus \{0\}$, satisfying $\int \min(1, \|v\|^2)d\mu(v) < \infty$ and

$$\sum_{i=1}^{n} \mathbb{P}\{Z_{ni}^{\delta} \in E\} \to \Phi|_{\{\|v\|>\delta\}}(E) \tag{2.5a}$$

for every $\delta > 0$ and every Borel set E satisfying $\Phi(\partial E) = 0$.

(ii) Q is a nonnegative definite quadratic form with

$$Q(u) = \lim_{\delta\to 0} \overline{\lim} \operatorname{Var}\left(\sum_{i=1}^{n} \langle Z_{ni\delta}, u\rangle\right) \tag{2.5a}$$

for each $u \in \mathbb{R}^p$ where $\overline{\lim}$ means either $\underline{\lim}$ or $\overline{\lim}$ as $n \to \infty$.

The centering constants a_{ni}, a may be chosen to be

$$a_{ni} = \mathbb{E}Z_{ni\delta_0}; \qquad\qquad a := \int_{\{\delta_0<\|x\|\leq 1\}} x d\Phi \tag{2.5c}$$

for any $0 < \delta_0 < 1$ such that $\Phi\{\|v\| = \delta_0\} = 0$.

Proof. It follows from condition (ii) and Lindeberg's CLT (see e.g. [15]) that if $\delta_n \to 0$ slowly enough, then

$$\sum_{i=1}^{n} (Z_{ni\delta_n} - \mathbb{E}Z_{ni\delta_n}) \rightsquigarrow N(0, Q) \tag{2.6}$$

We will now show convergence of $\sum_{i=1}^{n}(Z_{ni}^{\delta} - \mathbb{E}Z_{ni}^{\delta})$ for every continuity value δ, i.e. those for which $\Phi\{\|v\| = \delta\} = 0$. From Theorem 2.2.6 ($\mathcal{B}$ is the collection of Borel sets in $\mathbb{R}^p$),

$$\sup_{E\in\mathcal{B}} \left| \mathcal{L}\left(\sum_{i=1}^{n} Z_{ni}^{\delta}\right)(E) - \text{Pois}\left(\sum_{i=1}^{n} \mathcal{L}(Z_{ni}^{\delta})\right)(E) \right| \leq \sum_{i=1}^{n} (\mathbb{P}\{Z_{ni} > \delta\})^2$$

But, since hypothesis (i) holds and Z_{ni} is uniformly infinitesimal,

$$\sum_{i=1}^{n} (\mathbb{P}\{Z_{ni} > \delta\})^2 \leq \left(\sup_{i\leq n} \mathbb{P}\{Z_{ni} > \delta\}\right) \sum_{i=1}^{n} \mathbb{P}\{Z_{ni} > \delta\} \to 0$$

as $n \to \infty$. Hence for each δ, we have an infinitely divisible surrogate (usually called the accompanying law), $\text{Pois}(\sum_{i=1}^{n} \mathcal{L}(Z_{ni}^{\delta}))$, to approximate the distribution of the sum $\sum_{i=1}^{n} Z_{ni}^{\delta}$. Fortunately, infinitely divisible distributions are well understood. By Lemma 2.2.4(i),

$$\begin{aligned}
&\log \text{ch.f.Pois}\left(\sum_{i=1}^{n} \mathcal{L}(Z_{ni}^{\delta})\right) - i\left\langle u, \sum_{i=1}^{n} a_{ni} \right\rangle \\
=&\sum_{i=1}^{n} \int \left(e^{i\langle u,x\rangle} - 1 - i\langle u, x\rangle \mathbb{1}\{\|x\| \leq \delta_0\}\right) d\mathcal{L}(Z_{ni}^{\delta})(x) \\
\to &\int \left(e^{i\langle u,x\rangle} - 1 - i\langle u, x\rangle \mathbb{1}\{\|x\| \leq \delta_0\}\right) d\Phi|_{\{\|v\|>\delta\}}(x)
\end{aligned}$$

with convergence holding by hypothesis (i) and the Portmanteau Theorem. Finally

as $\delta \to 0$, we use Theorem 1.1.2 to get

$$\int_{\|v\|>\delta} \left(e^{i\langle u,x\rangle} - 1 - i\langle u,x\rangle \mathbb{1}\{\|x\| \le \delta_0\}\right) d\Phi(x) \to i\langle u,a\rangle + \log \text{ch.f.Pois}(\Phi)$$

The proof will be completed by the decoupage de Lévy in Lemma 2.2.1 with $E_n = \{\|v\| > \delta_n\}$ and $W_{1i} \sim W_{2i}$ mean $\mathbb{P}\{Z_{ni\delta_n} \in E_n\}$ Bernoulli RV's if

$$\sum_{i=1}^{n} (W_{1i} - W_{2i}) Z^{\flat}_{ni\delta_n} \rightsquigarrow 0$$

Indeed, we even have convergence in L^2:

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{n} (W_{1i} - W_{2i}) Z^{\flat}_{ni\delta_n}\right)^2 =& \sum_{i=1}^{n} \mathbb{E}(W_{1i} - W_{2i})^2 \mathbb{E} Z^{\flat 2}_{ni\delta_n} \\
=& 2\sum_{i=1}^{n} \mathbb{E}(W_{1i}) \mathbb{E}(1 - W_{2i}) \mathbb{E} Z^{2}_{ni\delta_n} \\
\le& 2 \max_{i\le n} (\mathbb{E} Z^{\flat 2}_{ni\delta_n}) \sum_{i=1}^{n} \mathbb{P}\{\|Z_{ni}\| > \delta_n\} \\
\to& 0
\end{aligned}$$

for $\delta_n \to 0$ slowly enough. □

2.3 Applications to LASSO

Apply Theorem 2.2.7 to the sequence of random variables $Z_{ni} := b_n^{-1} \xi_i x_i$ in $\mathbb{R}^p$. Because x_i are fixed, the distributions of Z_{ni} are supported by the countable union of lines

$$\bigcup_{i=1}^{\infty} x_i \cdot (-\infty, \infty),$$

Interestingly, the limit distribution of Z_{ni} will seldom have the same support. The following lemma allows us to utilize the support of Z_{ni} with no structural assumptions

on the limit. For any continuous function $\Gamma(\theta) : \mathbb{S}^{p-1} \to [0, \infty]$, call $\Gamma(\theta)$ the polar representation of $E = \{u \in \mathbb{R}^p : \|u\| \leq \Gamma(u/\|u\|)\}$.

Lemma 2.3.1. *Let Φ_n be a sequence of finite measures on $\mathbb{R}^p$ and Π be the collection of subsets of $\mathbb{R}^p$ with a polar representation. If there is a measure Φ such that $\Phi_n(E^c) \to \Phi(E^c)$ for all $E \in \Pi$, then $\Phi_n \rightsquigarrow \Phi$.*

Proof. Π is closed under finite unions and intersections ($\max(\Gamma_1, \Gamma_2)$ and $\min(\Gamma_1, \Gamma_2)$ are continuous if Γ_1, Γ_2 are), and

$$(E_1 \setminus E_2) \cap (F_1 \setminus F_2) = (E_1 \cap F_1) \setminus (E_2 \cup F_2)$$

so

$$\Pi \setminus \Pi := \{E_1 \setminus E_2 : E_1, E_2 \in \Pi\}$$

is a π-system (nonempty and closed under finite intersections).

Next, for any nonzero $w \in \mathbb{R}^p$, write B_w (resp. $\text{int}(B_w)$) for the closed (resp. open) ball centered at w of radius less than $\|w\|$ and write B_0 for a closed ball centered at 0 of radius exactly $\|w\|$. Set

$$\begin{aligned} E_1 =& B_0 \cup B_w \\ E_2 =& B_0 \setminus \text{int}(B_w) \end{aligned}$$

To see that $E_1 \in \Pi$, note that the length of a tangent segment from the origin to B_w is less than $\|w\|$. Hence, the point of tangency lies inside B_0. Think of E_2 as the death star, with an OSHA non-compliant hypermatter reactor core. For the purposes of this thesis however, we mostly care about the crater looking thing. Accordingly, $w \in E_1 \setminus E_2 \subset B_w$. Theorem 2.2 in [24] then says that if $\Phi_n/|\Phi_n|$ is a sequence of

probability measures, then

$$\left(\frac{\Phi_n}{|\Phi_n|} \rightsquigarrow \frac{\Phi}{|\Phi|}\right) \quad \text{iff} \quad \left(\frac{\Phi_n}{|\Phi_n|}(E) \rightsquigarrow \frac{\Phi}{|\Phi|}(E) \ \text{ for all } E \in \Pi\right)$$

to jump from probability measures to finite measures as the theorem claims, we only need that $\Phi_n(\mathbb{R}^p) \to \Phi(\mathbb{R}^p)$. This is true since $\mathbb{R}^p \in \Pi$. □

Now, we state the only condition needed for convergence of Z_{ni}.

(F1) There is a finite Borel measure φ on the sphere $\mathbb{S}^{p-1} \subset \mathbb{R}^p$ defined by

$$\varphi(E) := \lim_{n\to\infty} \sum_{i=1}^{n} \left(r^{-\alpha}(t_{ni}) \mathbb{1}\left\{\frac{x_i}{\|x_i\|} \in -E\right\} + R^{-\alpha}(t_{ni}) \mathbb{1}\left\{\frac{x_i}{\|x_i\|} \in E\right\}\right)$$

for all Borel sets E satisfying $\varphi(\partial E) = 0$ and where $t_{ni} := b_n/\|x_i\|$.

Theorem 2.3.2. *Assume (F1). Let $\mu[a,\infty) = a^{-1/\alpha}$ and $\Phi\{da, d\theta\} = \mu\{da\}\varphi\{d\theta\}$. Then, we have convergence in distribution*

$$Z_n := \sum_{i=1}^{n} b_n^{-1} \xi_i x_i \rightsquigarrow Z_\infty \sim [0, 0, \Phi]$$

Proof. Again, let $Z_{ni} = b_n^{-1}\xi_i x_i$. We verify conditions (2.5a-2.5c) of Theorem 2.2.7 in order.

Fix a truncation level $\delta > 0$. By the portmanteau theorem for finite measures and Lemma 2.3.1, we only need to check (2.5a) on E^c for every E with a polar decomposition and $\Phi|_{\{\|v\|>\delta\}}(\partial E^c) = 0$. Meanwhile, each of the supports of $Z_{ni\delta}$ are contained in the countable union of rays $\bigcup_{i=1}^{\infty} x_i \cdot ((-\infty, -\delta] \cup [\delta, \infty))$. Hence, to each E^c are associated sequences $a_i^-, a_i^+ \geq \delta$ such that

$$E^c \cap \left(\bigcup_{i=1}^{\infty} x_i \cdot ((-\infty, -\delta] \cup [\delta, \infty)) \right) = \bigcup_{i=1}^{\infty} x_i \cdot ((-\infty, -a_i^-] \cup [a_i^+, \infty))$$

Φ-almost everywhere. For uniqueness of representation, assume that $a_i^+ = a_j^+$ (resp. $a_i^+ = a_j^-$) and $a_i^- = a_j^-$ (resp. $a_i^- = a_j^+$) when x_i is a positive (resp. negative) multiple of x_j. We will check convergence for sets E^c where $E \in \Pi$. Call $t_{ni} := b_n / \|x_i\|$.

$$\begin{aligned}
\sum_{i=1}^{n} \mathbb{P}\{Z_{ni} \in E^c\} &= \sum_{i=1}^{n} \mathbb{P}\{-\xi_i \geq a_i^- t_{ni}\} + \mathbb{P}\{\xi_i \geq a_i^+ t_{ni}\} \\
&= \sum_{i=1}^{n} r^{-\alpha}(a_i^- t_{ni}) + R^{-\alpha}(a_i^+ t_{ni})
\end{aligned}$$

Now since $E \in \Pi$, it has a polar decomposition Γ_E. By the Uniform Convergence Theorem 1.1.5 and the Times One Lemma 1.4.1, since $-a_i^-, a_i^+ \geq \delta$ and $\min_{i \leq n} t_{ni} \to \infty$, we have

$$\begin{aligned}
\sum_{i=1}^{n} \mathbb{P}\{Z_{ni} \in E^c\} &= \sum_{i=1}^{n} (a_i^-)^{-\alpha} r^{-\alpha}(t_{ni}) + (a_i^+)^{-\alpha} R^{-\alpha}(t_{ni}) \\
&= \sum_{i=1}^{n} \left(\Gamma_E \left(-\frac{x_i}{\|x_i\|} \right) \right)^{-\alpha} r^{-\alpha}(t_{ni}) + \left(\Gamma_E \left(\frac{x_i}{\|x_i\|} \right) \right)^{-\alpha} R^{-\alpha}(t_{ni}) \\
&\to \int \Gamma_E(\theta)^{-\alpha} \varphi\{d\theta\} \\
&= \int_{E^c} \Phi\{dr, d\theta\}
\end{aligned}$$

and (2.5a) holds. Next is the Gaussian part. Since the Lévy representation in Theorem 1.1.2 is unique, we can call $Q(t)$ the Gaussian component. In our case, where ξ_i are i.i.d., limiting distributions are always stable and have no Gaussian component, so we show (2.5a) for $Q(t) = 0$. Again, let $t_{ni} = b_n / \|x_i\|$, $\delta > 0$, and $\|u\| = 1$,

$$\begin{aligned}
\mathrm{Var}\langle Z_{ni}\mathbb{1}\{\|Z_{ni}\|<\delta\},u\rangle \leq & \mathbb{E}\langle Z_{ni}\mathbb{1}\{\|Z_{ni}\|<\delta\},u\rangle^2 \\
\leq & \mathbb{E}[\ \|Z_{ni}\|^2\mathbb{1}\{\|Z_{ni}\|<\delta\}\ ] \\
\leq & \int_0^{\delta^2}\mathbb{P}\{\|Z_{ni}\|^2>\theta\}d\theta \\
= & \int_0^{\delta}\mathbb{P}\left\{\|Z_{ni}\|^2>\theta^2\right\}\cdot 2\theta d\theta \\
= & \int_0^{\delta}\mathbb{P}\left\{|\xi_i|>\theta t_{ni}\right\}\cdot 2\theta d\theta \\
= & \int_0^{\delta t_{ni}}\mathbb{P}\left\{|\xi_i|>\theta\right\}\cdot\frac{2\theta d\theta}{t_{ni}^2}
\end{aligned}$$

Denote the regularly varying function of order $-\alpha$

$$\mathcal{R}^{-\alpha}(t):=\mathbb{P}\left\{|\xi_i|>\theta\right\}=r^{-\alpha}(\theta)+R^{-\alpha}(\theta)$$

Karamata's Theorem 1.1.10 says

$$\frac{\int_0^{\delta t_{ni}}\theta\mathcal{R}^{-\alpha}(\theta)d\theta}{(\delta t_{ni})^2\mathcal{R}^{-\alpha}(\delta t_{ni})}\to\frac{1}{2-\alpha}$$

as $t_{ni}\to\infty$ and regular variation of $\mathcal{R}^{-\alpha}(\theta)$ says

$$\frac{\mathcal{R}^{-\alpha}(\delta t_{ni})}{\mathcal{R}^{-\alpha}(t_{ni})}\to\delta^{-\alpha}$$

as $t_{ni}\to\infty$. By the Times One Lemma 1.4.1, we can approximate summing over

$i \le n$. For some constant $c > 1$ and large n,

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=1}^{n}\langle Z_{ni}\mathbb{1}\{\|Z_{ni}\| < \delta\}, u\rangle\right) \le & \frac{2c}{2-\alpha}\sum_{i=1}^{n}\frac{\mathcal{R}^{-\alpha}(t_{ni})(\delta t_{ni})^{2-\alpha}}{t_{ni}^{2-\alpha}} \\ = & \frac{2c\delta^{2-\alpha}}{2-\alpha}\sum_{i=1}^{n}\mathcal{R}^{-\alpha}(t_{ni}) \\ \to & \frac{2c\delta^{2-\alpha}}{2-\alpha}|\varphi|\end{aligned}$$

as $n \to \infty$. Letting $\delta \to 0$ shows that $Q(u) = 0$ for all $u \in \mathbb{R}^p$.

Finally, with $\mathbb{E}|\xi_i| < \infty$, (1.11) implies condition (2.5c) of Theorem 2.2.7. □

Theorem 2.3.2 is stated with exact knowledge of the distribution of ξ_i, but we would like to reduce the requirement. The following conditions are "nearly" equivalent in that the regularly varying parts come from the random noise. It is usually preferable to turn this into a condition involving data. So, we introduce the following conditions:

(F1′) There is a finite Borel measure φ_0 on the sphere $\mathbb{S}^{p-1} \subset \mathbb{R}^p$ defined by

$$\varphi_0(E) := \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left(\|x_i\|^{\alpha}\left((1-d)\mathbb{1}\left\{\frac{x_i}{\|x_i\|} \in -E\right\} + d\mathbb{1}\left\{\frac{x_i}{\|x_i\|} \in E\right\}\right)\right)$$

for all Borel sets E satisfying $\varphi_0(\partial E) = 0$. ($d$ comes from **F3′**)

(F2′) There is $0 < \kappa < \alpha$ such that

$$\sup_n \frac{1}{n}\sum_{i=1}^{n}\|x_i\|^{\alpha+\kappa} < \infty$$

(F3′) As $t \to \infty$, with $\mathcal{R}^{-\alpha}(t) = r^{-\alpha}(t) + R^{-\alpha}(t)$,

$$\frac{\mathbb{P}\{\xi_i > t\}}{\mathbb{P}\{|\xi_i| > t\}} = \frac{R^{-\alpha}(t)}{\mathcal{R}^{-\alpha}(t)} \to d \in [0,1]$$

The purpose of the next corollary is to replace the annoying $\mathcal{R}^{-\alpha}(t_{ni})$ with $\frac{\|x_i\|^\alpha}{n}$.

Corollary 2.3.3. *Assume (F1′), (F2′), (F3′). Let* $\mu[r,\infty) = r^{-1/\alpha}$, *and* $\Phi_0\{dr, d\theta\} = \mu\{dr\}\varphi_0\{d\theta\}$. *Then, we have the convergence in distribution*

$$Z_n := \sum_{i=1}^{n} b_n \xi_i x_i \rightsquigarrow Z_\infty \sim [0, 0, \Phi_0]$$

Proof. We only need to show that φ_0 satisfies (F1) under the new hypotheses. Denote $\mathcal{R}^{-\alpha}(t) = R^{-\alpha}(t) + r^{-\alpha}(t)$. Simply note that

$$\left| \frac{R^{-\alpha}(t)}{\mathcal{R}^{-\alpha}(t)} - d \right| \to 0$$

by (F3′). Again, set $t_{ni} = b_n/\|x_i\|$. The Times One Lemma along with the combination of condition (F2′) and the Plus Zero Lemma 1.4.2 give

$$\begin{aligned}\lim_{n\to\infty} \sum_{i=1}^{n} \left(R^{-\alpha}(t_{ni}) \mathbb{1}\left\{ \frac{x_i}{\|x_i\|} \in E \right\} \right) &= \lim_{n\to\infty} \sum_{i=1}^{n} \left(d(\mathcal{R}^{-\alpha}(t_{ni})) \mathbb{1}\left\{ \frac{x_i}{\|x_i\|} \in E \right\} \right) \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(\|x_i\|^\alpha \cdot d\mathbb{1}\left\{ \frac{x_i}{\|x_i\|} \in E \right\} \right)\end{aligned}$$

Similarly,

$$\lim_{n\to\infty} \sum_{i=1}^{n} \left(r^{-\alpha}(t_{ni}) \mathbb{1}\left\{ \frac{x_i}{\|x_i\|} \in -E \right\} \right) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(\|x_i\|^\alpha \cdot (1-d) \mathbb{1}\left\{ \frac{x_i}{\|x_i\|} \in -E \right\} \right)$$

We conclude that $\varphi(E) = \varphi_0(E)$ for all Borel sets E satisfying $\varphi_0(\partial E) = 0$. $\square$

Remark 2.3.4. *The proof of Theorem 2.3.2 still holds if the covariates x_i change with n, since Theorem 2.2.7 is formulated for triangular arrays.*

Now, asymptotics for the minimizers $\hat{u}$ follow easily. We introduce two more conditions to handle the rest of the objective in the LASSO criterion.

(F2) $b_n^{-2} X^T X =: C_n \to C_\infty \succ 0$ element-wise (using $\succ$ to mean positive definite).

(F3) $\lambda_n / b_n \to \lambda_\infty > 0$

Lemma 2.3.5. *Assume (F2)-(F3), and either (F1) or each of (F1′), (F2′) and (F3′). Then*

$$\hat{u}_n := b_n^{-1}(\hat{\beta}_n - \beta) = O_P(1)$$

Proof. Let N_1 be large enough so that $C_\infty \succ 0$ renders

$$b_n^{-2}\|Xu\|_2^2 = \langle u, C_n u\rangle \geq \frac{1}{2}\langle u, C_\infty u\rangle \geq \frac{\eta}{2}\|u\|_1^2 \tag{2.7}$$

for all $n \geq N_1$ and $\eta > 0$ the minimum eigenvalue of C_∞. In the same fashion, let N_2 be large enough that $\lambda \leq 2\lambda_\infty$ for all $n \geq N_2$. From (2.7) and the Basic Inequality (1.9-BI),

$$\frac{\eta}{2}\|\hat{u}_n\|_1^2 - 2\|\hat{u}_n\|\|Z_n\| - 2\lambda_\infty\sqrt{p}\|\hat{u}_n\| \leq 0$$

Dividing through by $\|\hat{u}_n\|$, we see that

$$\|\hat{u}_n\| \leq \frac{4}{\eta}(\|Z_n\| + \lambda_\infty\sqrt{p}) \tag{2.8}$$

for all $n \geq \max(N_1, N_2)$. Then, Theorem 2.3.2 or Corollary 2.3.3 imply that $\|Z_n\| = O_p(1)$. □

Theorem 2.3 of [10] follows from the basic inequality (1.9-BI) and the Marcinkiewicz SLLN:

Remark 2.3.6 (Strong Consistency)**.** *Suppose ξ_i is centered as $\mathbb{E}\xi_i = 0$ with an α-moment, and assume $\sup_{i\geq 1} \|x_i\| < \infty$. Then, under conditions (F2) and (F3), $\hat{u}_n = O(1)$ almost surely.*

Proof. Lemma 3.2 of [10] is the weighted Marcinkiewicz SLLN, which shows $\|Z_n\| = o(1)$ a.s. The inequality in (2.8) shows that $\|\hat{u}_n\| = O(1)$ a.s. and that as remarked in [10], the rate is slowed from $o(1)$ a.s. for $\lambda_n = 0$ (OLS) to $O(1)$ a.s. under (F3). □

Theorem 2.3.7. *Under (F1)-(F3),*

$$\hat{u}_n \rightsquigarrow \arg\min_{u\in\mathbb{R}^p} V_\infty(u)$$

where

$$V_\infty(u) := \langle u, C_\infty u\rangle - 2\langle u, Z_\infty\rangle - 2\lambda_\infty \sum_{j=1}^{p} \begin{cases} u_j \mathrm{Sign}\beta_j & \text{if } \beta_j \neq 0 \\ |u_j| & \text{if } \beta_j = 0 \end{cases},$$

Z_∞ is the limit as in Theorem 2.3.2, and $C \to C_\infty \succ 0$ elementwise. Refer to (1.7a-1.7c) for definitions.

Proof. Lemma 2.3.5 establishes $\hat{u}_n$ as uniformly tight, i.e. for every $\epsilon > 0$, there exists a compact interval $K \subset \mathbb{R}$ such that $\mathbb{P}\{\|\hat{u}_n\| \notin K\} < \epsilon$ for large n, and so, $\mathcal{E}_n(\hat{u}) = o_p(1)$. Theorem 2.3.2 gives $\langle u, Z_n\rangle \rightsquigarrow \langle u, Z_\infty\rangle$ in $\ell^\infty(K)$, and $\langle u, C_n u\rangle \to$

$\langle u, C_\infty u\rangle$ in $\ell^\infty(K)$ since pointwise convergence implies uniform convergence over finite sets

$$\max_{1\le i,j\le p} |(C_n)_{ij} - (C_\infty)_{ij}| \to 0$$

Hence, with $V_n(u)$ from (1.8), $V_n(u) \rightsquigarrow V_\infty(u)$ in $\ell^\infty(K)$ for every compact $K \subset \mathbb{R}^p$. By strict convexity, $V_\infty(u)$ has a unique minimum. The proof is finished by Theorem 1.6.1. □

3. VARIABLE DESIGN, FIXED NUMBER OF REGRESSORS

In this section, we think of the design X as coming from a class $\mathfrak{X}$ and we think of quantities from section 2 as functions of $X \in \mathfrak{X}$. For example, if $\Psi \in \mathfrak{X}$ has columns ψ_i, then $Z_n(\Psi) := n^{-1/\alpha} \sum_{i=1}^n \psi_i \xi_i$. We discuss the centered scaled LASSO $\hat{u}_n(X)$ as a random element of $\ell^\infty(\mathfrak{X}, \mathbb{R}^p)$, the space of maps $\mathfrak{X} \to \mathbb{R}^p$ with finite norm $\|u\|_{\mathfrak{X}} := \sup_{X \in \mathfrak{X}} \|u(X)\|$. As is usual, define the norm on $\ell^\infty(\mathfrak{X})$ by

$$|v|_{\mathfrak{X}} := \sup_{X \in \mathfrak{X}} |v(X)|$$

and $e_i \in \ell^\infty(\mathfrak{X}, \mathbb{R}^p), e_{ij} \in C(\mathfrak{X})$ to be coordinates:

$$e_i(X) := x_i$$

$$e_{ij}(X) := x_{ij}$$

3.1 A General CLT

Theorem 4.2 from [1] gives sufficient conditions for triangular arrays of random $\ell^\infty(\mathfrak{X})$-valued elements to converge to an infinitely divisible limit with no Gaussian component. We paraphrase a special case of this theorem, specifically for parameters which fit their Example 4.1(1) ($\varphi(x) = x^{1-1/\alpha}$) and for Borel measurable processes that concentrate on a separable subspace of $\ell^\infty(\mathfrak{X})$.

Theorem 3.1.1. *Suppose $(Z_{ni})_{i \le n}$ is a row-wise independent triangular array of random elements in $\ell^\infty(\mathfrak{X})$. Suppose also that $(Z_{ni})_{i \le n}$ concentrates on a separable*

subspace $\mathcal{V} \subset \ell^\infty(\mathfrak{X})$ *with envelope*

$$F(v) := \sup_{X \in \mathfrak{X}} |v(X)|$$

so that $F(v) < \infty$ *for all* $v \in \mathcal{V}$. *Assume the following*

(i) For every $\epsilon > 0$,

$$\sup_n \sum_{i=1}^n \mathbb{P}\{|Z_{ni}|_{\mathfrak{X}} > \epsilon\} < \infty$$

(ii) For every $\epsilon > 0$, *there is a compact (convex, symmetric)* $K \subset \ell^\infty(\mathfrak{X})$ *s.t.*

$$\overline{\lim} \sum_{i=1}^n \mathbb{P}\{Z_{ni}(\cdot) \notin K, |Z_{ni}|_{\mathfrak{X}} > \epsilon\} < \epsilon \tag{3.1}$$

(iii) There is a semimetric d *on* $\mathfrak{X}$ *and a Borel probability measure* μ *on* $(\mathfrak{X}, d)$ *such that*

(a) $\lim_{\epsilon \to 0} \sup_{X \in \mathfrak{X}} \int_0^\epsilon (-\ln \mu(B_d(X,t)))^{1-1/\alpha} dt = 0$ *with* $\sup_{X \in \mathfrak{X}}$ *finite for* $\epsilon = \infty$, *and*

(b) There are constants $\sigma > 0, n_0 > 0$, *and* $L_1 \geq 1$ *such that for all* $\Psi \in \mathfrak{X}, l \geq L_1, n \geq n_0$, *and* $\delta > 0$,

$$\sum_{i=1}^n \mathbb{P}\left\{ \sup_{X \in B_d(\Psi, \delta)} |Z_{ni}(X) - Z_{ni}(\Psi)| > \sigma \delta / l^{1/\alpha} \right\} \leq l/3 \tag{3.2}$$

(iv) For each k *and* $X_1, \ldots X_k \in \mathfrak{X}$, *the triangular array of* $\mathbb{R}^k$*-valued random vectors* $(Z_{ni}(X^1), \ldots, Z_{ni}(X^k))_{i \leq n}$ *is infinitesimal and the sequence*

$$\left\{ \sum_{i=1}^n (Z_{ni}(X^1), \ldots, Z_{ni}(X^k)) \right\}$$

converges in law to an infinitely divisible law with a degenerate Gaussian component.

Then, the sequence of random elements $\{\sum_{i=1}^{n} Z_{ni}(X) : X \in \mathfrak{X}\}$ *converges in law to a Radon infinitely divisible measure on* $\ell^\infty(\mathfrak{X})$ *with a degenerate Gaussian component. The finite dimensional distributions are given by (iv).*

3.2 Choosing a Semimetric

By Prohorov's Theorem, for a sequence of processes $\mathbb{M}_n$ to converge, it must be uniformly tight and have a unique limit. Uniqueness follows if the finite dimensional distributions converge, to which Theorem 2.2.7 applies. Usually, proving tightness is the tougher task. The next criterion allows us to jump from finite subsets of an index set to totally bounded subsets (under a certain semimetric). Note, asymptotic tightness is equivalent to uniform tightness in separable, completely metrizable spaces (visit Theorem 1.5.7 of [35] for reference).

Theorem 3.2.1. *A sequence of processes* $\mathbb{M}_n$ *in* $\ell^\infty(\mathfrak{X})$ *is asymptotically tight iff there is a semimetric* ρ *on* $\mathfrak{X}$ *which makes* $\mathfrak{X}$ *totally bounded and* $\mathbb{M}_n$ *asymptotically uniformly* ρ*-equicontinuous in probability, i.e. for all* $\epsilon > 0$*,*

$$\lim_{\delta \to 0} \overline{\lim} \, \mathbb{P} \left\{ \sup_{\rho(X,\Psi)<\delta} |\mathbb{M}_n(X) - \mathbb{M}_n(\Psi)| > \epsilon \right\} = 0 \tag{3.3}$$

Ideally, $\rho(X, \Psi)$ should follow $|\mathbb{M}_n(X) - \mathbb{M}_n(\Psi)|$ closely, although $\mathbb{M}_n$ is random. If first moments are finite, it would be correct to let $\rho(X, \Psi) = \overline{\lim} \, \mathbb{E}|\mathbb{M}_n(X) - \mathbb{M}_n(\Psi)|$. But, since our random variables lie in the domain of attraction of a stable, we have a quicker route. As n gets large, only the tails of ξ_i are important, which we know to vary regularly of order α. For each $j \leq p$, we create a seminorm $|\cdot|_j$ by comparing the distribution of ξ_i to a symmetric strictly α-stable (SαS) distribution, which also

has regularly varying tails of order α and sums nicely, according to (1.2). Indeed, if ξ_i were SαS, and $\mathbb{M}_{nj}(X) = n^{-1/\alpha}\sum_{i=1}^{n} x_{ij}\xi_i$, we get

$$\begin{aligned}\mathbb{M}_{nj}(X) - \mathbb{M}_{nj}(\Psi) =& n^{-1/\alpha}\sum_{i=1}^{n}(x_{ij}-\psi_{ij})\xi_i \\ &\sim \left(\frac{1}{n}\sum_{i=1}^{n}|x_{ij}-\psi_{ij}|^{\alpha}\right)^{1/\alpha}\xi_1\end{aligned}$$

by equation (1.2) with $1 < \alpha < 2$. Thus, a natural choice for ρ is $\rho(X,\Psi) := \sup_j |X-\Psi|_j$ where

$$|X|_j := \overline{\lim}\,\frac{1}{n}\sum_{i=1}^{n}|x_{ij}|^{\alpha} \tag{3.4}$$

3.3 Convergence of the Cross Term

Let us make (F1′) and (F2′) uniform over $\mathfrak{X}$.

(D1) There is a finite Borel measure $\varphi^{\mathfrak{X}}$ on the $\|\cdot\|_{\mathfrak{X}}$-sphere in $\ell^{\infty}(\mathfrak{X},\mathbb{R}^p)$ so that

$$\varphi^{\mathfrak{X}}(E) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(1-d)\,\|e_i\|_{\mathfrak{X}}^{\alpha}\,\mathbb{1}\left\{\frac{e_i}{\|e_i\|_{\mathfrak{X}}}\in -E\right\} + d\,\|e_i\|_{\mathfrak{X}}^{\alpha}\,\mathbb{1}\left\{\frac{e_i}{\|e_i\|_{\mathfrak{X}}}\in E\right\}$$

for all Borel sets E satisfying $\varphi^{\mathfrak{X}}(\partial E) = 0$. Again, denote by $|\varphi^X|$ the total variation of φ^X.

(D2) For each j, there is a κ such that

$$\sup_n \frac{1}{n}\sum_{i=1}^{n}\|e_i\|_{\mathfrak{X}}^{\alpha+\kappa} < \infty$$

It follows from $|\varphi^{\mathfrak{X}}| < \infty$ that $|\cdot|_j$ is finite on $\mathfrak{X}$ for every $j \in \mathbb{N}$. In view of Theorem 3.1.1, we state the following conditions:

(D3) For every $\epsilon > 0$, there is $I_\epsilon \subset \mathbb{N}$ and $(q_i)_{i\in I_\epsilon}$ such that

$$\sup_{\substack{\rho(X,0)\leq\delta \\ i\in I_\epsilon^c}} q_i\|x_i\| \to 0$$

as $\delta \to 0$ and

$$\overline{\lim}\, \frac{1}{n}\sum_{I_\epsilon} q_i^{-\alpha} < \epsilon \qquad \text{and} \qquad \overline{\lim}\, \frac{1}{n}\sum_{I_\epsilon^c} |e_{ij}|_{\mathfrak{X}}^{\alpha} < \epsilon^{1+\alpha}$$

(D4) There is a Borel probability measure μ on $\mathfrak{X}$ satisfying

$$\lim_{\epsilon\to 0}\sup_{X\in\mathfrak{X}}\int_0^\epsilon (-\log \mu(B_\rho(X;t)))^{1/\alpha} dt = 0$$

with $\sup_{X\in\mathfrak{X}}$ finite for $\epsilon = \infty$.

Theorem 3.3.1. *Let $\mu[r,\infty) = r^{-1/\alpha}$ and $\varphi^{\mathfrak{X}}$ as in (D1) and*

$$\phi^{\mathfrak{X}}\{dr, d\theta\} = \mu\{dr\}\varphi^{\mathfrak{X}}\{d\theta\}$$

Then, under (D1)-(D4), (F3′), and the uniform version of (1.10), i.e.

$$\sup_{i\leq n}\|e_i\|_{\mathfrak{X}} = \sup_{\substack{i\leq n \\ X\in\mathfrak{X}}}\|x_i\| = o\left(b_n\right) \tag{3.5}$$

we have that

$$Z_n := b_n^{-1}\sum_{i=1}^n e_i\xi_i$$

is tight as an element of $\ell^\infty(\mathfrak{X},\mathbb{R}^p)$. Hence,

$$Z_n \rightsquigarrow Z_\infty \sim [0,0,\phi^{\mathfrak{X}}]$$

Proof. Let $Z_{ni} = b_n^{-1} e_i \xi_i$ be regarded as a function $\mathfrak{X} \times \{1, \ldots, p\} \to \mathbb{R}$, i.e. $Z_{ni}(X, j) = b_n^{-1} x_{ij} \xi_i$. We will use Theorem 3.1.1 to prove that $\sum_{i=1}^n Z_{ni}$ is tight as a sequence of elements in $\ell^\infty(\mathfrak{X} \times \{1, \ldots, p\})$. First, condition (i) comes by

$$\begin{aligned}
\sum_{i=1}^n \mathbb{P}\{\|Z_{ni}\|_{\mathfrak{X}} > t\} &= \sum_{i=1}^n \mathbb{P}\left\{|\xi_i| > \frac{b_n t}{\|e_i\|_{\mathfrak{X}}}\right\} \\
&\to t^{-\alpha} |\varphi^{\mathfrak{X}}| \qquad \text{as} \quad n \to \infty \\
&< \infty
\end{aligned}$$

for every $t > 0$ (convergence holding thanks to equation (3.5) and the Times One Lemma 1.4.1).

As for (ii) of Theorem 3.1.1, let $\epsilon > 0$ and q, I_ϵ be as in (D3). Let $K_q := \text{conv.hull}\{a e_i : i \in I_\epsilon, |a| \leq q_i\}$. Therefore, if $X, \Psi \in K_q$, then $X - \Psi \in K_q$. Now, we use Ascoli's Theorem to prove K_q is compact. K_q is (uniformly) equicontinuous by

$$\begin{aligned}
\sup_{\substack{\rho(X,\Psi) \leq \delta \\ X, \Psi \in K_q}} \sup_{\substack{|a| \leq q_i \\ i \in I_\epsilon}} \|a e_i(X) - a e_i(\Psi)\| &= \sup_{\rho(X,\Psi) \leq \delta} \sup_{i \in I_\epsilon} q_i \|x_i - \psi_i\| \\
&= \sup_{\rho(X,0) \leq \delta} \sup_{i \in I_\epsilon} q_i \|x_i\| \\
&\to 0 \qquad (3.6)
\end{aligned}$$

as $\delta \to 0$, which holds by (D3). Then, pointwise boundedness holds if for each $X \in K_q$, we have

$$\sup_{\substack{|a| \leq q_i \\ i \in I_\epsilon}} \|a e_i(X)\| = \sup_{i \in I_\epsilon} q_i \|x_i\| < \infty$$

By (3.6), there must be $\delta > 0$ so that

$$\sup_{\rho(X,0)\leq\delta} \sup_{i\in I_\epsilon} q_i \|x_i\| < \infty$$

Since $\rho\left(\frac{\delta X}{\rho(X,0)}, 0\right) = \delta$, it must be that

$$\sup_{i\in I_\epsilon} q_i \left\| \frac{\delta x_i}{\rho(X,0)} \right\| < \infty$$

Also, ρ is finite by the comment after definition (3.4). It follows that $K_{qj} \subset \ell^\infty(\mathfrak{X})$ is equicontinuous and pointwise bounded, hence compact. Then, condition (ii) is met by Lemma 1.4.2 and (D3):

$$\begin{aligned}
\overline{\lim} \sum_{i=1}^n \mathbb{P}\{Z_{ni} \notin K_q, \|Z_{ni}\|_{\mathfrak{X}} > \epsilon\} =& \overline{\lim} \sum_{i=1}^n \mathbb{P}\left\{ |\xi_i| > \max\left(b_n q_i, \frac{\epsilon b_n}{\|e_i\|_{\mathfrak{X}}} \right) \right\} \\
\leq& \overline{\lim}\ \frac{p}{n} \sum_{i\in I_\epsilon} q_i^{-\alpha} + \overline{\lim}\ \frac{p\epsilon^{-\alpha}}{n} \sum_{i\notin I_\epsilon} \|e_i\|_{\mathfrak{X}}^\alpha \\
\leq& 2p\epsilon |\varphi^{\mathfrak{X}}|
\end{aligned}$$

Condition (iiia) is the same as (D4). For condition (iiib) use (D1) with $\sigma = 3|\varphi^{\mathfrak{X}}|^{1/\alpha}$

$$\begin{aligned}
& \sum_{i=1}^n \mathbb{P}\left\{ \sup_{\rho(X,\Psi)<\delta} \|Z_{ni}(X) - Z_{ni}(\Psi)\| > 3\delta \left(\frac{|\varphi^{\mathfrak{X}}|}{l} \right)^{1/\alpha} \right\} \\
=& \sum_{i=1}^n \mathbb{P}\left\{ \sup_{\rho(X,0)<\delta} \|b_n^{-1} x_i \xi_i\| > 3\delta \left(\frac{|\varphi^{\mathfrak{X}}|}{l} \right)^{1/\alpha} \right\} \\
\to& \left(3\left(|\varphi^{\mathfrak{X}}|/l\right)^{1/\alpha}\right)^{-\alpha} \cdot \frac{1}{n} \sum_{i=1}^n \sup_{\rho(X,0)<1} \|x_i\|^\alpha \\
\leq& l/3^\alpha
\end{aligned}$$

Choose N large enough that this convergence is within $1/3 - 1/3^\alpha$ for $n \geq N$.

Finally, to show that the finite dimensional distributions converge to stable limits, let $X^1, \cdots X^k \in \mathfrak{X}$ be arbitrary and consider the process $Z^*_{ni} := b_n^{-1} e_i \xi_i$ indexed by $\mathfrak{X}^* := \{0 =: X^0, X^1, \ldots, X^k \in \mathfrak{X}\}$. By the Cramer-Wold device and linearity of Z^*_{ni}, convergence as in $Z^*_{ni} \rightsquigarrow Z^*_\infty \sim [0, 0, \varphi^{\mathfrak{X}^*}]$ is equivalent to the convergence

$$Z^*_{ni}\left(\sum_{i=1}^{n} c_l X^l\right) \rightsquigarrow Z^*_\infty \left(\sum_{l=0}^{k} c_l X^l\right)$$

Or, assuming $\mathfrak{X}$ to be convex and symmetric, we only need pointwise convergence. Since $\varphi^{\mathfrak{X}}|_X$ coincides with φ from (F1′) and conditions (D1),(D2) are stronger than (F1′), (F2′), we are done.

□

3.4 LASSO

Now, we get to throw in the rest of the objective V_n^X. Assume the following.

(D5) For each $X \in \mathfrak{X}$, there exists a $p \times p$ matrix $C_\infty(X) \succ 0$ so that

(a) With the semimetric ρ from (3.4) and the operator norm of $\mathbb{R}^p \to \mathbb{R}^p$ denoted by $\|\cdot\|_{\text{op}}$,

$$\overline{\lim} \sup_{\rho(X,\Psi)<\delta} \|C_\infty^X - C_\infty^\Psi\|_{\text{op}} \to 0$$

as $\delta \to 0$

(b) $\inf_{X \in \mathfrak{X}} \|C_\infty^X\|_{\text{op}} \geq \eta_{\min} > 0$ and

$$\sup_{X \in \mathfrak{X}} \|C_n(X) - C_\infty(X)\|_{\text{op}} \to 0$$

(D6) $\lambda_n / b_n \to \lambda_\infty$

Theorem 3.4.1. *(D5b),(D6) and* $|Z_n|_{\mathfrak{X}} = O_P(1)$ *imply*

$$\|\hat{u}_n(X)\| \lesssim \|Z_n(X)\| + \lambda_\infty$$

for each $X \in \mathfrak{X}$*, up to a constant not depending on* X*. Hence*

$$\sup_{X \in \mathfrak{X}} \|\hat{u}_n(X)\| = O_P(1)$$

Proof. Take n large enough that $\sup_{X\in\mathfrak{X}} \|C_n^X - C_\infty^X\|_{\text{op}} \leq \frac{1}{2} \inf_{X\in\mathfrak{X}} \|C_\infty(X)\|_{\text{op}}$ and $\lambda_n/b_n \leq 2\lambda_\infty$. Then,

$$\begin{aligned} 0 \geq \inf_{u\in\mathbb{R}^p} V_n(X;u) &\geq \langle \hat{u}_n(X), C_n(X)\hat{u}_n(X)\rangle - 2\|\hat{u}_n(X)\|\|Z_n(X)\| - 2\lambda_\infty\|u\| \\ &\geq \frac{\eta_{\min}}{2}\|\hat{u}_n(X)\|^2 - 2\|\hat{u}_n(X)\|(\|Z_n(X)\| + \lambda_\infty) \\ &= \left(\frac{\eta_{\min}}{2}\|\hat{u}(X)\| - 2(\|Z_n(X)\| + \lambda_\infty)\right)\|\hat{u}^X\|_1 \end{aligned}$$

and so

$$\|\hat{u}_n(X)\| \leq \frac{4}{\eta_{\min}}(\|Z_n(X)\| + \lambda_\infty)$$

Take $\sup_{\mathfrak{X}}$.

□

Next is a partial converse to Theorem 3.4.1, taking standard bounded eigenvalues.

Theorem 3.4.2. *Assume (D5), (D6) and* $\sup_{\mathfrak{X}} \|C_n(X)\|_{\text{op}} \leq \eta_{\max} < \infty$ *for all* n*. Then, for any* $X \in \mathfrak{X}$*,*

$$\|Z_n(X)\| - 3\lambda_\infty \lesssim \|\hat{u}_n(X)\|$$

for each $X \in \mathfrak{X}$, up to a constant not depending on X. Hence

$$\|\hat{u}_n\|_{\mathfrak{X}} = O_P(1) \Longrightarrow \|Z_n\|_{\mathfrak{X}} = O_P(1)$$

Proof. Pick $\|\dot{u}(X)\| = 1$ so that $\langle \dot{u}(X), Z_n(X) \rangle = \|Z_n(X)\|$. Then, as long as $\lambda_n / b_n \leq 2\lambda_\infty$, we have for each $X \in \mathfrak{X}$

$$\begin{aligned}
\inf_{u \in \mathbb{R}^p} V_n(X; u) &\leq \inf_{t \in \mathbb{R}} V_n(X; \dot{u}(X)t) \\
&\leq \inf_{t \in \mathbb{R}} \left(\eta_{\max} t^2 - 2t\|Z_n(X)\| + 2\lambda_\infty t \right) \\
&= -\frac{(\|Z_n(X)\| - \lambda_\infty)^2}{\eta_{\max}}
\end{aligned}$$

Also,

$$\begin{aligned}
\inf_{u \in \mathbb{R}^p} V_n(X; u) &\geq -2\|\hat{u}_n(X)\|\|Z_n(X)\| - 2\lambda_\infty \|\hat{u}_n(X)\|_1 \\
&= -2\|\hat{u}_n(X)\|(\|Z_n(X)\| + \lambda_\infty)
\end{aligned}$$

Combining these,

$$\begin{aligned}
\|\hat{u}_n(X)\| &\geq \frac{(\|Z_n(X)\| - \lambda_\infty)^2}{2\eta_{\max}(\|Z_n(X)\| + \lambda_\infty)} \\
&= \frac{1}{2\eta_{\max}} \left(\|Z_n(X)\| - 3\lambda_\infty + \frac{4\lambda_\infty^2}{\|Z_n(X)\| + \lambda_\infty} \right)
\end{aligned}$$

□

To continue, keep a fixed design $\Psi = [\psi_1 \dots \psi_p]$ for some $\psi_j \in \mathbb{R}^p$. We want to analyze $\hat{u}^X$ for X relative to Ψ, so we center accordingly.

Lemma 3.4.3. *Under (D1)-(D6), we have for every $\epsilon > 0$, as $\delta \to 0$*

$$\overline{\lim}\, \mathbb{P}\left\{ \sup_{\substack{\rho(X,\Psi)|<\delta \\ \|u\|\leq 1}} |V_n(X;u) - V_n(\Psi;u)| > \epsilon \right\} \to 0$$

Proof. First, rewrite

$$V_n(X;u) - V_n(\Psi;u) = \langle u, (C_n(X) - C_n(\Psi))u\rangle - 2\langle u, Z_n(X) - Z_n(\Psi)\rangle$$

which splits the proof into

$$\lim_{\delta\to 0} \overline{\lim}\, \mathbb{P}\left\{ \sup_{\rho(X,\Psi)<\delta} \|Z^X - Z^\Psi\| > \epsilon \right\} \to 0$$

(true by Theorems 3.3.1 and 3.2.1) and

$$\begin{aligned}
\lim_{\delta\to 0} \overline{\lim} \sup_{\substack{\rho(X,\Psi)<\delta \\ \|u\|\leq 1}} |\langle u, (C_n(X) - C_n(\Psi))u\rangle| &\leq \lim_{\delta\to 0} \overline{\lim} \sup_{\rho(X,\Psi)<\delta} \|C_n(X) - C_n(\Psi)\| \\
&\leq \lim_{\delta\to 0} \overline{\lim} \sup_{\rho(X,\Psi)<\delta} \|C_n(X) - C_\infty(X)\| \\
&\quad + \lim_{\delta\to 0} \overline{\lim} \sup_{\rho(X,\Psi)<\delta} \|C_\infty(X) - C_\infty(\Psi)\| \\
&\quad + \lim_{\delta\to 0} \overline{\lim} \sup_{\rho(X,\Psi)<\delta} \|C_\infty(\Psi) - C_n(\Psi)\| \\
&\to 0
\end{aligned}$$

which is true by (D5). □

Now we're ready for a CLT.

Theorem 3.4.4. *If $\mathfrak{X}$ is totally bounded under ρ, then (D1)-(D6) imply that $\hat{u}_n(X)$ is uniformly tight as a random element in $\ell^\infty(\mathfrak{X})$, so*

$$\hat{u}_n(X) \rightsquigarrow \underset{u\in\mathbb{R}^p}{\arg\min} \left[\langle u, C_\infty(X)u\rangle - 2\langle u, Z_\infty(X)\rangle + \lambda_\infty \sum_{j=1}^{p} \left\{ \begin{array}{ll} u_j \text{Sign}\beta_j & \text{if} \quad \beta_j \neq 0 \\ |u_j| & \text{if} \quad \beta_j = 0 \end{array} \right\} \right]$$

uniformly in $\mathfrak{X}$. Recall, $Z_\infty(X)$ is the w-limit as in Theorem 3.3.1.

Proof. From Theorem 3.4.1, $\mathcal{E}(u) = o_P(1)$. Also, $\langle u, C_n(X)u\rangle \to \langle u, C_\infty(X)u\rangle$ in $\ell^\infty(\mathfrak{X})$ by (D5a), $Z_n(X) \rightsquigarrow Z_\infty(X)$ in $\ell^\infty(\mathfrak{X})$ by Theorem 3.3.1, and $\lambda_n/b_n \to \lambda_\infty$ by (D6). If follows that

$$V_n^X(u) \rightsquigarrow V_\infty^X(u)$$

in $\ell^\infty(\mathfrak{X})$. To apply Theorem 1.6.1, we only need to prove $\hat{u}^X$ is uniformly tight, or since $\mathfrak{X}$ is ρ-totally bounded,

$$\overline{\lim}\ \mathbb{P}\left\{ \sup_{\rho(X,\Psi)<\delta} \|\hat{u}^X - \hat{u}^\Psi\| > \epsilon \right\} \to 0 \tag{3.7}$$

for every $\epsilon > 0$, as $\delta \to 0$.

To this end, notice for each $\|u\|, \|w\| \leq 1$ and $X, \Psi \in \mathfrak{X}$,

$$V_n(X;u) - V_n(X;w) + 2\max_{\{u,w\}} |V_n(X;\cdot) - V_n(\Psi;\cdot)| > V_n(\Psi;u) - V_n(\Psi;w)$$

Then, set $u = \hat{u}_n(X)$ and $w = \hat{u}_n(\Psi)$. Pick n large enough that $\Psi^T\Psi \succ 0$, so for

every $\epsilon > 0$ there is $\tilde{\epsilon}$ s.t.

$$\begin{aligned}\|\hat{u}_n(X) - \hat{u}_n(\Psi)\| > \epsilon \Rightarrow & V_n(\Psi; \hat{u}_n(X)) - V_n(\Psi; \hat{u}_n(\Psi)) > \tilde{\epsilon} \\ \Rightarrow & 2 \sup_{\|u\| \leq 1} |V_n(X; u) - V_n(\Psi; u)| > \tilde{\epsilon}\end{aligned}$$

Finally,

$$\overline{\lim}\ \mathbb{P}\left\{\sup_{\rho(X,\Psi)<\delta} \|\hat{u}_n(X) - \hat{u}_n(\Psi)\| > \epsilon\right\} \leq \overline{\lim}\ \mathbb{P}\left\{\sup_{\substack{\rho(X,\Psi)<\delta \\ \|u\| \leq 1}} |V_n(X; u) - V_n^(\Psi; u)| > \tilde{\epsilon}\right\}$$

3.7 then follows by Lemma 3.4.3

□

4. STOCHASTIC BOUNDEDNESS WITH INCREASING NUMBER OF REGRESSORS

When the number of regressors p is fixed as the number of data n grows, we can wait until the design becomes sufficiently uncorrelated (unless there is some inherent dependence in the designs). If, on the other hand, p is allowed to increase with n (as will be the case throughout this section, and sometimes emphasized by the subscript p_n), correlation among the regressors (columns of X) becomes a bigger problem. M-estimation as $p \to \infty$, even with nongaussian tails, has been studied in the 70's and 80's ([19, 27]).

No matter how fast p grows, solving for the LASSO solution is always a finite dimensional problem. Yet, concerning asymptotics, finite dimensional tools necessarily turn up lacking. We will think of β as a sequence of real numbers, and set $\beta_n = (\beta_j)_{j \le p_n}$. The regularization term suggests we measure the accuracy of the LASSO with $\|\hat{\beta}_n - \beta_n\|_1$, which in turn suggests that x_i be thought as an element of the dual $(\ell^1)^* = \ell^\infty$. Then, $Z_n := b_n^{-1} \sum_{i=1}^n x_i \xi_i$ will be a random process indexed by the unit ball in ℓ^1, denoted by B_1. Note, when we treat B_1 as an index set, we will use a seminorm other than $\|\cdot\|_1$ that is tailored to fit with X.

4.1 Boundedness of Z_n

Now, we state Proposition 4.3 from [1] in the ℓ^∞ case, using constants as in their Example 4.1(1) ($H(x) = x^{1/\alpha}$):

Theorem 4.1.1. *Suppose $\{Z_{ni}\}_{i \le n}$ is a triangular array of row-wise independent ℓ^∞-valued random variables and $\{\epsilon_i\}_{i \ge 1}$ be a Rademacher sequence independent of $\{Z_{ni}\}_{i \le n}$. Assume the following*

(i)

$$\lim_{M\to\infty}\sup_n\sum_{i=1}^n \mathbb{P}\{\|Z_{ni}\|_\infty > M\} = 0$$

(ii) There is a seminorm $|\cdot|$ on the pointset B_1, and a probability measure μ on the $|\cdot|$-Borel sets satisfying

(a)

$$\lim_{\epsilon\to 0}\sup_{\|u\|_1\le 1}\int_0^\epsilon (-\ln\mu(B_{|\cdot|}(u,t)))^{1-1/\alpha}dt = 0$$

with $\sup_{\|u\|_1\le 1}$ *finite for* $\epsilon = \infty$*, and*

(b) There are constants $\sigma > 0, n_0 > 0$, and $L_1 \ge 1$ such that for all $\|u_1\| \le 1$, $l \ge L_1, n \ge n_0$, we have

$$\sum_{i=1}^n \mathbb{P}\left\{\sup_{|u|\le 1}|\langle u, Z_{ni}\rangle| > \sigma l^{-1/\alpha}\right\} \le l/3 \tag{4.1}$$

Similar to (3.4), we address (ii) by defining the seminorm on ℓ_1 as

$$|u|^\alpha := \overline{\lim}\,\frac{1}{n}\sum_{i=1}^n |\langle u, x_i\rangle|^\alpha$$

Then, we state a couple assumptions.

(P1) For some $\kappa > 0$,

$$d := \sup_n \sum_{i=1}^n \frac{\|x_i\|_\infty^{\alpha+\kappa}}{n} < \infty$$

(P2) There is a probability measure μ on the $|\cdot|$-Borel sets s.t.

$$\lim_{\epsilon\to 0}\sup_{\|u\|_1\le 1}\int_0^\epsilon (-\ln\mu(B_{|\cdot|}(u;t)))^{1-1/\alpha}dt = 0$$

with $\sup_{\|u\|\le 1}$ finite for $\epsilon = \infty$.

Then, we conclude that $\|Z_n\|_\infty$ is stochastically bounded:

Theorem 4.1.2. *Assume (P1), (P2). Then,* $\|b_n^{-1}\sum_{i=1}^n x_i\xi_i\|_\infty = O_p(1)$.

Proof. Let $Z_{ni} := b_n^{-1}x_i\xi_i$. According to Theorem 4.1.1, we only need the following in addition to (P2):

(i)

$$\begin{aligned}\sup_n \sum_{i=1}^n \mathbb{P}\{\|Z_{ni}\|_\infty > M\} &= \sup_n \sum_{i=1}^n \mathbb{P}\{|\xi_i| > Mb_n/\|x_i\|_\infty\} \\ &= \sup_n \sum_{i=1}^n \mathcal{R}^{-\alpha}\left(\frac{Mb_n}{\|x_i\|_\infty}\right) \\ &= \sup_n M^{-\alpha}\sum_{i=1}^n \mathcal{R}^{-\alpha}\left(\frac{b_n}{\|x_i\|_\infty}\right) + o(1) \\ &= \sup_n \frac{M^{-\alpha}}{n}\sum_{i=1}^n \|x_i\|_\infty^\alpha + o(1)\end{aligned}$$

by the Uniform Convergence Theorem 1.1.5 and Lemma 1.4.2. By (P1), the last line goes to 0 as $M \to \infty$.

(ii) For constants $\sigma = (4d)^{1/\alpha}$, $n_0 \geq 1$, and $l_0 \geq 1$ s.t. for all $\|u\|_1 \leq 1$, $l > l_0$,

$n > n_0$, and $\epsilon > 0$,

$$\begin{aligned}
\sum_{i=1}^{n} \mathbb{P}\{\|Z_{ni}\|_\infty > \sigma l^{-1/\alpha}\} &= \sum_{i=1}^{n} \mathbb{P}\left\{|\xi_i| > \frac{\sigma l^{-1/\alpha} b_n}{\|x_i\|_\infty}\right\} \\
&= \sum_{i=1}^{n} \mathcal{R}^{-\alpha}\left(\frac{\sigma l^{-1/\alpha} b_n}{\|x_i\|_\infty}\right) \\
&= \left(\sigma l^{-1/\alpha}\right)^{-\alpha} \sum_{i=1}^{n} \mathcal{R}^{-\alpha}\left(\frac{b_n}{\|x_i\|_\infty}\right) + o(1) \\
&= \frac{l}{n\sigma^\alpha} \sum_{i=1}^{n} \frac{l \cdot \|x_i\|^\alpha}{\sigma^\alpha} + o(1) \\
&\leq l/4 + o(1)
\end{aligned}$$

□

4.2 The Oracle

In the case that Rank $X < p$, it could be that $X\beta = 0$, in which case it is not possible to learn about β from our model. To begin to absolve this issue, define $S := \{j \leq p : \beta_j \neq 0\}$ (which depends on n) and the "restricted set" as $\mathcal{U} = \{u \in \mathbb{R}^p : \|u_{S^c}\|_1 \leq 4\|u\|_1\}$. Then, let us give some moduli of continuity, or compatibility conditions (so called for making different norms compatible, see [34]):

(P3) There is a constant $\zeta > 0$ such that

$$\inf_n \inf_{u \in \mathcal{U}} \frac{\|Xu\|}{b_n \|u_S\|_1} \geq \zeta > 0 \tag{4.2}$$

(P4) There is a constant $\eta > 0$ such that

$$\sup_n \sup_{u \in \mathbb{R}^S} \frac{\|Xu_S\|_\infty}{b_n \|u_S\|_1} \leq \eta$$

This is all we need before we use arguments similar to [6] in the proof of Theorem 4.2.1. For now, to gain perspective, suppose an oracle told us the support S of β beforehand. Then, there would be no reason to shrink the estimates, and we could regress X_S onto β with ordinary least squares as follows. Call

$$\tilde{\beta} = (X_S^T X_S)^{-1} X_S^T Y$$

which is a random element in $\mathbb{R}^S$. Substitute $Y = X\beta + \xi = X_S\beta_S + \xi$,

$$\begin{aligned}\tilde{\beta} =&(X_S^T X_S)^{-1} X_S^T (X_S\beta_S + \xi)\\ =&\beta_S + (X_S^T X_S)^{-1} X_S^T \xi\end{aligned}$$

Now, subtract β_S and left-multply by $C_S := b_n^{-2}(X_S^T X_S)$,

$$\begin{aligned}C_S(\tilde{\beta} - \beta_S) =&b_n^{-1}(b_n^{-1} X_S^T \xi)\\ =&b_n^{-1} Z_S\end{aligned} \tag{4.3}$$

From equation (4.3), we can get an upper bound for $\|\tilde{\beta} - \beta_S\|_1$ by assuming (P3), and a lower bound by assuming (P4). Let us start with the upper bound. Left-multiply equation (4.3) by $\tilde{\beta} - \beta_S$,

$$\langle \tilde{\beta} - \beta_S, C_S(\tilde{\beta} - \beta_S)\rangle = b_n^{-1}\langle \tilde{\beta} - \beta_S, Z_S\rangle$$

Now, $\tilde{\beta} - \beta_S \in \mathcal{R}$ because $\tilde{\beta}$ has support in S, so (P3) implies

$$\begin{aligned}\zeta^2\|\tilde{\beta} - \beta_S\|_1^2 \leq & b_n^{-1}\|\tilde{\beta} - \beta_S\|_1\|Z_S\|_\infty \\ \zeta^2\|\tilde{\beta} - \beta_S\|_1 \leq & b_n^{-1}\|Z_S\|_\infty \end{aligned} \tag{4.4}$$

Next, the lower bound will follow from (P4). Again, from equation (4.3),

$$\begin{aligned} b_n^{-1}\|Z_S\|_\infty &= \|C_S(\tilde{\beta} - \beta_S)\|_\infty \\ &\leq \eta\|\tilde{\beta} - \beta_S\|_1 \end{aligned} \tag{4.5}$$

Thus, with the addition of condition (P4), the bound in equation (4.4) is optimal up to a constant factor. Supposing condition (F1) for the design X_S, we have $\|Z_S\|_\infty = O_p(1)$, and so these bounds imply that $\|\tilde{\beta} - \beta\|_1 = O_p(b_n^{-1})$. Similar to Theorem 6.2 of [6] and Theorem 3 of [38], the next theorem shows that the LASSO $\hat{\beta}$ performs nearly as well as this "oracle" rate, which is the rate from using OLS regression on the support of β.

Theorem 4.2.1. *Suppose $\lambda_n/b_n \to \infty$ in addition to (P1)-(P3). It follows that $\|\hat{\beta}_n - \beta\|_1 = O_p(\lambda_n/b_n^2)$.*

Proof. Define $S := \{j \leq p : \beta_j \neq 0\}$ with $|S| = s$. Note that since $\lambda_n/b_n \to \infty$, Theorem 4.1.2 implies that

$$\|X^T\xi\|_\infty = b_n\|Z_n\|_\infty = O_p(b_n) = o_p(\lambda_n) \tag{4.6}$$

Consider $u \in \ell_1^p$ (e.g. $u = \hat{\beta}_n - \beta$) satisfying the Basic Inequality

$$\|Xu\|^2 - 2\langle Xu, \xi\rangle + \lambda_n\|u + \beta\|_1 \leq \lambda_n\|\beta\|_1 \tag{4.7-BI}$$

We will use the hypotheses to get an upper bound on u. Consider the case that $\|X^T\xi\|_\infty \leq \lambda_n/4$, which holds with high probability by (4.6). Use Hölder's inequality on equation (4.7-BI) to get

$$\|Xu\|^2 - 2\left(\frac{\lambda_n}{4}\right)\|u\|_1 + \lambda_n\|u+\beta\|_1 \leq \lambda_n\|\beta\|_1$$

Now, separate the S^c terms and the S terms, use reverse triangle inequality, then add $\lambda_n\|u_S\|_1/2$ to both sides

$$\begin{aligned}
\|Xu\|^2 - \frac{\lambda_n}{2}\|u_{S^c}\|_1 + \lambda_n\|u_{S^c}\|_1 &\leq \frac{\lambda_n}{2}\|u_S\|_1 + \lambda_n(\|\beta\|_1 - \|u_S+\beta\|_1) \\
\|Xu\|^2 + \frac{\lambda_n}{2}\|u_{S^c}\|_1 &\leq \frac{3\lambda_n}{2}\|u_S\|_1 \\
\|Xu\|^2 + \frac{\lambda_n}{2}\|u\|_1 &\leq 2\lambda_n\|u_S\|_1
\end{aligned}$$

This implies u belongs to the restricted set $\mathcal{U} = \{u \in \mathbb{R}^p : \|u_{S^c}\|_1 \leq 4\|u\|_1\}$. Hence, we can continue with

$$\|Xu\|^2 + \frac{\lambda_n}{2}\|u\|_1 \leq \ 2\lambda_n\left(\frac{\|Xu\|}{\zeta b_n}\right)$$

which immediately gives $\|Xu\| \leq 2b_n^{-1}\lambda_n/\zeta$. Turn the inequality on itself to rid u from the right hand side,

$$\|Xu\|^2 + \frac{\lambda_n}{2}\|u\|_1 \leq \ \frac{4\lambda_n^2}{\zeta^2 b_n^2}$$

Thus, we have a bound for both the prediction error $\|Xu\|$ and the estimation error $\|u\|_1$ of those u that satisfy (4.7-BI), including $u = \hat{\beta}_n - \beta$. □

5. INCREASING NUMBER OF REGRESSORS

Section 4 got close to the "exact" rate of convergence for the LASSO estimator when p is allowed to increase with n. Nevertheless, even if we set the noise $\xi = 0$, and $\lambda_n > 0$, LASSO will incur some amount of shrinkage and fail to exactly recover $\beta \neq 0$ (see [4]). In the same vein, any deterministic choice for λ_n will not compare well to the cross term. This seems to result in either negligible shrinkage or too much bias. Luckily, the square root LASSO, written $\sqrt{\text{LASSO}}$, introduced in [4], automatically handles the penalty level λ_n and exactly recovers β when $\xi = 0$. We define it by

$$\hat{\beta}_{\text{SQ}} := \underset{u \in \mathbb{R}^p}{\arg\min} [\|Y - Xu\| + \lambda_n \|u\|_1] \tag{5.1}$$

The most striking benefit to omitting the square is that if we scale the data (X, Y) and the penalty level λ_n by a constant, then $\hat{\beta}_{\text{SQ}}$ stays fixed. This proportionality suggests that λ_n may be more easily decided for $\sqrt{\text{LASSO}}$ than for LASSO, as mentioned before. Interestingly, [4] found that $\sqrt{\text{LASSO}}$ performs well even with heteroscedastic errors. Though we've only allowed i.i.d. errors, heteroscedasticity appears to be a natural pursuit due to our use of triangular arrays. Additionally, [4] found a bound for $\hat{\beta}_{\text{SQ}} - \beta$ in the prediction norm for an infinite variance case (when ξ_i is distributed like the t-distribution with two degrees of freedom). For this section, for computation's sake, let us assume ξ_i are symmetric Pareto distributed $\mathbb{P}\{\pm\xi_i > t\} = t^{-\alpha}$ (when $t \geq 1$) and $b_n = n^{1/\alpha}$.

The first challenge in finding an exact rate of convergence is to find conditions on X (perhaps sparse conditions) that subdue the crazy. We may even think of a

different kind of sparsity wherein a vector is sparse if most of its coordinates are only very small (as compared to exactly zero)[38]. Then, of course, we must modify our restricted eigenvalues to the new kind of sparsity.

Moreover, we readily accept condition (P3) to handle the case that $X\beta \approx 0$. Also, we would like to use Theorem 3.1.1 to prove convergence of the cross term. Verily, conditions (i),(ii) of Theorem 3.1.1 are necessary and contained in condition (C1) in subsection 5.2. Next, (P2) implies condition (iii), and Theorem 2.3.2 handles convergence of finite dimensional distributions. So, assume that

$$\sum_{i=1}^{n} Z_{ni} = n^{-1/\alpha} \sum_{i=1}^{n} x_i \xi_i \rightsquigarrow Z_\infty$$

Finally, we normalize the covariates to $n^{-1/\alpha}\|x_i\|_\alpha = 1$ for each $i \leq n$. Alternatively we could use penalty loadings (as in [4]), or weights, inside the ℓ^1 term. We suggest considering $\lambda_j = \left(\frac{1}{n}\sum_{i=1}^{n} \|x_i\|_\infty^\alpha\right)^{1/\alpha}$, or perhaps functions thereof, since we are expecting the objective $V_n(u)$ to converge, in preparation for Theorem 1.6.1.

5.1 Reformulation of $\sqrt{\text{LASSO}}$

Write the $\sqrt{\text{LASSO}}$ objective in terms of a local ℓ^1 parameter and center at 0 just as in section 1. This entails substituting $u + \beta$ for u in the objective of (1.6) then subtracting

$$\begin{aligned} \|X(0+\beta) - Y\| + \|(0+\beta)\|_1 =& \|Xu - Y\| + \|\beta\|_1 \\ =& \|\xi\| + \|\beta\|_1 \end{aligned}$$

Thus, the objective that $n^{1/\alpha}(\hat{\beta}_{\text{SQ}} - \beta)$ minimizes becomes

$$\begin{aligned}
&\|X(n^{-1/\alpha}u + \beta) - Y\| - \|\xi\| + \|(n^{-1/\alpha}u + \beta)\|_1 - \|\beta\|_1 \\
=&\|n^{-1/\alpha}Xu - \xi\| - \|\xi\| + \sum_{j=1}^{p} \lambda_j(|n^{-1/\alpha}u_j + \beta| - |\beta|) \qquad (5.2)
\end{aligned}$$

Reformulating $|n^{-1/\alpha}u_j + \beta| - |\beta|$ works the same as in the logic succeeding equation (1.6). But, with the square root, $\|\xi\|$ doesn't cancel right away. We can use the "conjugate trick" as in

$$\begin{aligned}
\|n^{-1/\alpha}Xu - \xi\| - \|\xi\| =& \frac{\|n^{-1/\alpha}Xu - \xi\|^2 - \|\xi\|^2}{\|n^{-1/\alpha}Xu - \xi\| + \|\xi\|} \\
=& \frac{n^{-2/\alpha}\|Xu\|^2 - 2n^{-1/\alpha}\langle u, X^T\xi\rangle}{\|n^{-1/\alpha}Xu - \xi\| + \|\xi\|}
\end{aligned}$$

Then, we will assume that $n^{-2/\alpha}\|Xu\|^2$ converges uniformly to a quadratic form $C_\infty(u)$. If we further scale the parameter by substituting $n^{-1/\alpha}u$ for u, the ξ's in the denominator of the last display overtake $n^{-1/\alpha}Xu$. In other words,

$$\frac{2|\xi\|}{\|n^{-1/\alpha}Xu - \xi\| + \|\xi\|} \xrightarrow{p} 1 \qquad (5.3)$$

Thus, the objective for $\sqrt{\text{LASSO}}$. Finally, multiply the objective (5.2) through by $2\|\xi\|$. It follows that the only effective difference between LASSO and $\sqrt{\text{LASSO}}$ is that λ is multiplied by $2\|\xi\|$,

$$\hat{\beta}_{\text{SQ}} - \beta \approx \arg\min \ \langle u, C_n u\rangle - 2\langle u, Z_n\rangle + 2\lambda_n\|\xi\| \cdot \|u + \beta\|_1$$

Usually, the penalty is assumed to be of a larger order than Z_n (as was the case in

section 4). To find the exact rate of convergence, though, requires λ to be of the same order as Z_n. Ostensibly, the randomization of λ will facilitate this goal and manage the balance between the noise and the bias introduced by the regularization term.

5.2 Controlling the Cross Term

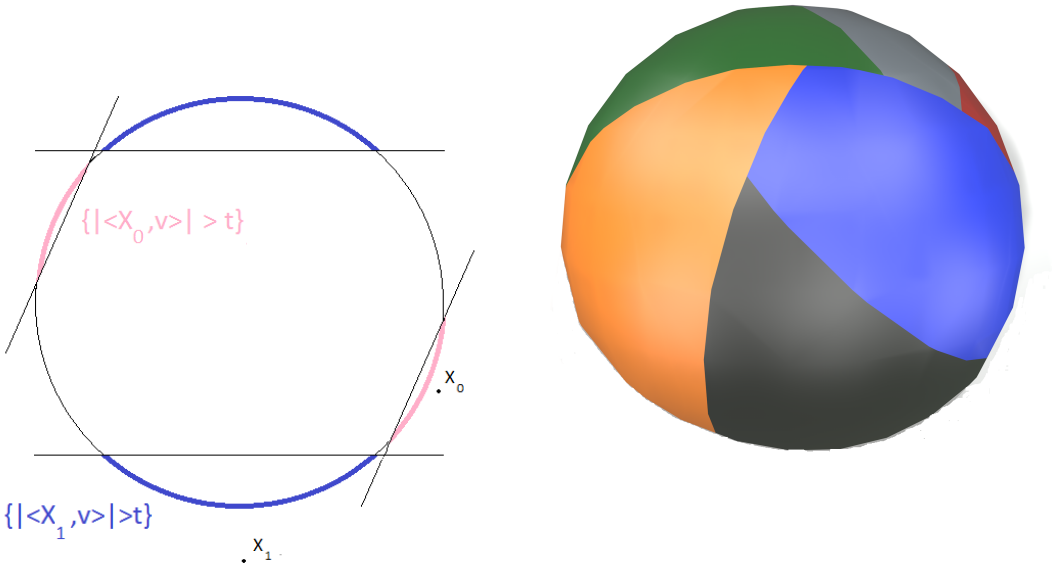
Since u is chosen *after* ξ is realized, u may align somewhat with $X^T\xi$ when minimizing the objective. So, we believe that Cauchy-Schwarz $|\langle Xu, \xi\rangle| \leq \|u\|_1 \|X^T\xi\|_\infty$ is not too loose.

Similar to section 4, we still want the penalization to be heavy enough to allow us to focus on a "restricted set." Recall how we normalized $\|X_j\|_\alpha = n^{1/\alpha}$. We should also divide through by $\|\xi\|$ (normalize $\|\xi\| = 1$). So, we are compelled to analyze the distributions of $\xi/\|\xi\|$ as random elements of $\mathbb{S}^{p-1}$. Since ξ_i are i.i.d., these distributions will be symmetric and we might as well analyze the order statistics $\xi_{(1)} \geq \xi_{(2)} \geq \cdots \geq \xi_{(n)}$. Interestingly, the Pareto distribution is characterized by the fact that if $\xi_1, \ldots, \xi_n$ are Pareto, then $\xi_{(1)}/\xi_{(N)}, \ldots, \xi_{(N-1)}/\xi_{(N)}$ are again Pareto distributed and independent of $\xi_{(N)}$. Of course, this observation suggests normalizing to $\|\xi\|_\infty = 1$. On the other hand, $\|\xi\|$ appears in the effectively random $\sqrt{\text{LASSO}}$ penalty, so normalizing to $\|\xi\| = 1$ makes sense, too. We've found the former to be useful for intuitively handling ξ, and the latter to be useful for formulating results.

Define

$$\begin{aligned} Q_n(t) :=& \mathbb{P}\left\{ n^{-1/\alpha} \|X^T\xi\|_\infty > t \cdot \frac{\|\xi\|}{n^{1/\alpha}} \right\} \\ =& \mathbb{P}\left\{ \left| \left\langle X_j, \frac{\xi}{\|\xi\|} \right\rangle \right| > t \text{ for some } j \leq p_n \right\} \end{aligned}$$

We seek conditions on X that promise $Q_n(\lambda_\infty) \to 0$ for some constant λ_∞. Now, we



(a) The $\mathbb{R}^n$ sphere ($n = 2$) with regions $\{v \in \mathbb{S}^1 : |\langle X_j, v\rangle| > t\}$ colored for $j = 0, 1$.

(b) The $\mathbb{R}^n$ sphere (n=3) partitioned according to which j maximizes $|\langle X_j, v\rangle|/\|X_j\|_\alpha$.

Figure 5.1: $\xi/\|\xi\|$ as a random map on the $\mathbb{R}^n$ sphere.

can think of $\xi/\|\xi\|$ as a random element of the sphere in $\mathbb{R}^n$.

Figure 5.1a illustrates a way of calculating $Q_n(t)$. Given a cutoff level t, each X_j will define a region $\{v : |\langle v, X_j\rangle| > t\}$ of the $\mathbb{R}^n$ sphere. The probability $Q_n(t)$ is the probability that $\xi/\|\xi\|$ belongs to the union of these regions. Analyzing the behavior of $Q_n(t)$ is crucial for us to utilize the $\sqrt{\text{LASSO}}$.

Thx Parth Sarin and Cinema4D

Figure 5.1b is related to Figure 5.1a. In it, each X_j also determines a region, but instead of comparing to a parameter t, it compares to all the other X_j's. Accordingly, define

$$P_j := \{v \in \mathbb{R}^n : \|v\| = 1, |\langle X_j, v\rangle| \geq |\langle X_k, v\rangle| \text{ for all } k \leq p_n\}$$

which will partition the sphere (almost everywhere). It gives an idea of who the key players are among the X_j's. The bigger (Lebesgue measure) regions tend to

correspond to X_j's that are more important than those corresponding to smaller regions. Also, since we normalize $\|X_j\|_\alpha = n^{1/\alpha}$ and $\|X_j\|_\alpha \approx \|X_j\|$ when X_j has a few coordinates much larger than the rest, the regions concentrating near an axis tend to correspond to more important X_j's. If X_j has many moderate coordinates and none extreme, then $\langle X_j, \xi/\|\xi\|\rangle$ tends to be only moderate. All the action seems to happen near the axes. Therefore, we guess that approximating $\xi/\|\xi\|$ by n equiprobable point masses at e_i will maintain sufficient conditions.

$$\begin{aligned} Q_n(t) \approx & \frac{1}{n}\#\{i \le n : |x_{ij}| > t \text{ for some } j \le p\} \\ = & \frac{1}{n}\#\{i \le n : \|x_i\|_\infty > t\} \end{aligned}$$

This approximation leads to

(C1) For every ϵ, there is an $M \ge 1$ s.t.

$$\sup_n \frac{1}{n}\sum_{i=M}^{n} \|x_i\|_\infty^\alpha < \epsilon$$

Another consideration is the growth of p. Authors such as Huber [19] and Portnoy [27, 28] have assumed $p = o(n^{1/2})$ in the case of finite variance errors. We think $p = o(n^{1/\alpha})$ is appropriate for errors with α tails. Rather, $\sqrt{\text{LASSO}}$ should allow p to increase even faster, as long as the cardinality of the support of β is $o(n^{1/\alpha})$.

Conjecture 5.2.1. *Assume (P1)-(P3), (C1) and $p = o(n^{1/\alpha})$. Let $\lambda_n \to \lambda_\infty$. Then, the following is sufficient for $\hat{\beta}_{\text{SQ}} - \beta = O_P(n^{-1/\alpha})$:*

$$\frac{1}{n}\#\{i \le n : \|x_i\|_\infty > \lambda_\infty\} \to 0$$

5.3 Subregressions

For tightness, we are looking to characterize compact sets (complete and totally bounded). In normed sequence spaces, relatively compact sets are those which are bounded and have uniformly small tails in norm. In ℓ^1, this amounts to a finite envelope. That is, each element $v \in \ell^1$ determines a compact set $\{u : |u_j| \leq |v_j| \text{ for all } j\}$, and every compact set in ℓ^1 is a subset of such an example.

$\hat{u}_n$ will be tight if for every ϵ, there is a compact $K \subset \ell^1$ such that $\mathbb{P}\{\hat{u}_n \notin K\} < \epsilon$. Since we are presuming at this stage that $\hat{u}_n$ is stochastically bounded, we can focus on proving $\hat{u}_n$ also has small tails. The best idea we have found is to perform a subregression, or a regression where we leave out finitely many coordinates $j = 1, \ldots, M$ and subtract off the projections of $X_{M+1}, \ldots X_p$ onto $\text{span}\{X_1, \ldots, X_p\}$. Then, if the norm of this estimate is small with high probability, we will have a compact set. It's beautiful. Unfortunately, I just attempted to cut this section with CTRL+X on a computer that has cleverly substituted the CTRL button for a FN button (leaving just 'x' and no undo).

6. CONCLUSIONS

The LASSO estimator is a very hot topic due to its quick computation time and its effectiveness for sparse data. Much has been written about LASSO under gaussian errors, or even errors with finite variance. Yet, little is known in the case that the errors have infinite variance.

In section 2, we explored the asymptotic behavior of LASSO in a basic multivariate setup with i.i.d. regularly varying errors. Our methods seem to generalize to nonidentical, infinitely divisible error distributions, but notation would of course bear the burden.

Section 3 extended that study to consider infinitely many data matrices at once. We found conditions that promised the same fidelity of estimates for each data matrix, simultaneously. This could potentially be useful for error-in-variables models, as long as noisy data matrices tend to fall in a class satisfying our conditions. Also, our result feels like a stability theorem of machine learning. That is, if we perturb the data matrix some, we can still expect LASSO to perform well.

A very interesting setup comes when we allow the number of variables to increase as we gather more data. Huber and Portnoy began studying such regressions, including asymptotics, in [19, 27, 28]. Today, asymptotics seem to be understudied, compared to inequalities like error bounds. We found some success with LASSO when we categorized the variables as OLD or NEW, then subtracted from NEW the projections of NEW onto OLD (representing the explanatory data not yet used in OLD), and performed another regression (which we called a subregression) on the modified NEW set.

REFERENCES

[1] Niels T Andersen, Evarist Giné, and Joel Zinn. The central limit theorem for empirical processes under local conditions: the case of radon infinitely divisible limits without gaussian component. Transactions of the American Mathematical Society, 308(2):603–635, 1988.

[2] Aloisio Araujo and Evarist Giné. The central limit theorem for real and Banach valued random variables. John Wiley & Sons, 1980.

[3] Sylvain Arlot, Alain Celisse, et al. A survey of cross-validation procedures for model selection. Statistics surveys, 4:40–79, 2010.

[4] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. Biometrika, 98(4):791–806, 2011.

[5] Dimitri P Bertsekas. Convex optimization theory. Athena Scientific Belmont, MA, 2009.

[6] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. The Annals of Statistics, pages 1705–1732, 2009.

[7] Nicholas H Bingham, Charles M Goldie, and Jef L Teugels. Regular variation, volume 27. Cambridge university press, 1989.

[8] Peter Bühlmann and Sara Van De Geer. Statistics for high-dimensional data: methods, theory and applications. Springer Science & Business Media, 2011.

[9] Emmanuel J Candès, Yaniv Plan, et al. Near-ideal model selection by 1 minimization. The Annals of Statistics, 37(5A):2145–2177, 2009.

[10] A. Chatterjee and S.N. Lahiri. Strong consistency of lasso estimators. Sankhya A, 73(1):55–78, 2011.

[11] Albert Cohen, Wolfgang Dahmen, and Ronald DeVore. Compressed sensing and best -term approximation. Journal of the American mathematical society, 22(1):211–231, 2009.

[12] David L Donoho and Xiaoming Huo. Uncertainty principles and ideal atomic decomposition. Information Theory, IEEE Transactions on, 47(7):2845–2862, 2001.

[13] RM Dudley and Marek Kanter. Zero-one laws for stable measures. Proceedings of the American Mathematical Society, pages 245–252, 1974.

[14] William Feller. An Introduction to Probability Theory and Its Applications: Volume 1. J. Wiley & sons, 1960.

[15] William Feller. An introduction to probability and its applications, vol. ii. Wiley, New York, 1971.

[16] Arie Feuer and Arkadi Nemirovski. On sparse representation in pairs of bases. IEEE Transactions on Information Theory, 49(6):1579–1581, 2003.

[17] Wayne A Fuller. Measurement error models, volume 305. John Wiley & Sons, 2009.

[18] Boris Vladimirovich Gnedenko, Andreĭ Nikolaevich Kolmogorov, Kai Lai Chung, Joseph L Doob, and Pao-Lu Hsu. Limit distributions for sums of independent random variables, volume 233. Addison-Wesley Reading, Massachusetts, 1968.

[19] Peter J Huber. Robust regression: asymptotics, conjectures and monte carlo. The Annals of Statistics, pages 799–821, 1973.

[20] Mikhail K Kozlov, Sergei P Tarasov, and Leonid G Khachiyan. The polynomial solvability of convex quadratic programming. USSR Computational Mathematics and Mathematical Physics, 20(5):223–228, 1980.

[21] Benoit Mandelbrot. The variation of certain speculative prices. Journal of Business, pages 294–419, 1963.

[22] Mark M Meerschaert and Hans-Peter Scheffler. Limit distributions for sums of independent random vectors: Heavy tails in theory and practice, volume 321. John Wiley & Sons, 2001.

[23] Balas Kausik Natarajan. Sparse approximate solutions to linear systems. SIAM journal on computing, 24(2):227–234, 1995.

[24] Billingsley Patrick. Convergence of probability measures. New York etc.: John Wiley and Sons, 1968.

[25] Valentin V Petrov. Limit theorems of probability theory, volume 358. Clarendon Press, Oxford, 1995.

[26] Siméon Denis Poisson. Probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilitiés. Paris, France: Bachelier, 1, 1837.

[27] Stephen Portnoy. Asymptotic behavior of m-estimators of p regression parameters when p2/n is large. i. consistency. The Annals of Statistics, pages 1298–1309, 1984.

[28] Stephen Portnoy. Asymptotic behavior of m estimators of p regression parameters when p2/n is large; ii. normal approximation. The Annals of Statistics, pages 1403–1417, 1985.

[29] EL Rvaceva. On domains of attraction of multi-dimensional distributions. Select. Transl. Math. Statist. and Probability, 2:183–205, 1961.

[30] Claude Elwood Shannon. Communication in the presence of noise. Proceedings of the IRE, 37(1):10–21, 1949.

[31] Steven Smith. Digital Signal Processing: A Practical Guide for Engineers and Scientists: A Practical Guide for Engineers and Scientists. Newnes, 2013.

[32] Philippe Soulier. Some applications of regular variation in probability and statistics. Escuela Venezolana de Matemáticas, 2009.

[33] Ryan J Tibshirani, Jonathan Taylor, et al. Degrees of freedom in lasso problems. The Annals of Statistics, 40(2):1198–1232, 2012.

[34] Sara van de Geer. The deterministic lasso. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) Zürich, 2007.

[35] Aad W Van Der Vaart and Jon A Wellner. Weak Convergence. Springer, 1996.

[36] B.L. van der Waerden. Mathematische Statistik. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2013.

[37] William J Welch. Algorithmic complexity: Three np-hard problems in computational statistics. Journal of Statistical Computation and Simulation, 15(1):17–25, 1982.

[38] Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. The Annals of Statistics, pages 1567–1594, 2008.